DOCUMENT RESUME

ED 348 398 TM 018 878

TITLE	Assessing Student Achievement in the States. The First Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1990 Trial State Assessment.
INSTITUTION	National Academy of Education, Stanford, Calif.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
PUB DATE	92
NOTE	128p.
AVAILABLE FROM	National Academy of Education, Stanford University, School of Education CERAS-507, Stanford, CA 94305-3084 ($20 prepaid).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Academic Achievement; Advisory Committees; *Educational Assessment; Educational Policy; Elementary Secondary Education; Federal Government; Government Role; Grade 8; *Guidelines; Mathematics Achievement; *National Programs; Policy Formation; Program Validation; State Programs; *State Surveys; *Student Evaluation
IDENTIFIERS	*Trial State Assessment (NAEP)

ABSTRACT

This report presents findings of the National Academy of Education Panel on the Evaluation of the Trial State Assessment (TSA) concerning the 1990 National Assessment of Educational Progress (NAEP) mathematics TSA program. The Panel's recommendations regarding the continuation of TSAs for review by the Congress, states, and Executive Branch of the Federal Government include: Congress should approve the continuation of state NAEPs; Congress should authorize additional TSAs; the prohibition against reporting NAEP data below the state level should continue; NAEP content frameworks should evolve toward increased testing of higher order thinking skills; continuity throughout the NAEP multiphase process should be assured; the random monitoring of schools' administration of the 1990 mathematics assessment should continue; future authorizations for state NAEPs should include adequate resources to sample private school students to increase the comparability of results from one state to another; beginning with the 1994 assessment, a national sample of out-of-school 17-year-olds should be assessed; the NCES should continue to provide the media with specific examples of interpretations of the results and construct new tables and graphs to effectively display the data. Seven tables, 6 figures, 37 references, and an appendix with synopses of studies for the 1990 evaluation of the NAEP TSA are included. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED348398

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

R. BENNET

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Assessing Student Achievement in the States

The First Report
of the National Academy of Education
Panel on the Evaluation of the NAEP Trial State Assessment:
1990 Trial State Assessment

Panel Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

TM018878

BEST COPY AVAILABLE

Assessing Student Achievement in the States

Copyright © 1992 by the National Academy of Education
All rights reserved.
Printed on recycled paper in the United States of America
Single copy (prepaid): $20
Studies for the Evaluation of the National Assessment of Educational Progress Trial State Assessment—1990: $10
National Academy of Education
Stanford University
School of Education CERAS-507
Stanford CA, 94305-3084
(415) 725-1003

Assessing Student Achievement in the States

The First Report
of the National Academy of Education
Panel on the Evaluation of the NAEP Trial State Assessment:
1990 Trial State Assessment

Panel Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

The National Academy of Education

The National Academy of Education is composed of scholars and education leaders who "promote scholarly inquiry and discussion concerning the ends and means of education, in all its forms, in the United States and abroad." Our current active membership is limited to seventy-five scholars. The heart of the Academy is found in the serious and vital exchanges that take place in our regular meetings and the special commissions we establish. Throughout our 26-year history, the Academy has been called upon by governmental and other agencies to conduct special studies and reviews in the public interest. The most comprehensive evaluation of the National Assessment of Educational Progress ever conducted was organized by the Academy and directed by Lamar Alexander and H. Thomas James in 1986. We not only served as fiscal agent and coordinator of the Alexander-James Panel, we also provided a thoroughly independent review of that endeavor. Most recently, we completed a study, funded by the Carnegie Corporation of New York, to recommend priorities for the funding of educational research entitled "Research and the Renewal of Education."

As the national trend toward education reform broadens and accelerates, we focus again on NAEP. This time we contribute the intellectual leadership and the coordination for an independent evaluation of the 1990 and 1992 Trial State Assessments, which provide state-by-state comparisons of educational achievement, the first such use of NAEP data. Specifically, we are organizing and conducting a carefully constructed portfolio of studies using a range of methodological approaches. A Panel of experts, selected by the Academy on the basis of their distinguished accomplishments, manages and coordinates the study portfolio. The Panel and the Academy examine a range of major issues: state-by-state and state-to-national comparisons; the validity and reliability of the assessment data; content, curriculum, and consensus processes; operations, sampling, data analysis; and the methods of reporting results of the assessment.

This is the first of two mandated reports; the second will be issued in 1993. A third and final capstone report will draw together information from various sources to document the quality and validity of the Trial State Assessment data.

This report carries the approval of the Panel and has been carefully reviewed and approved by the Academy's Executive Council, acting as a Committee of Readers. On behalf of the Executive Council of the Academy and its entire membership, I am pleased to present this report. Its findings and recommendations should assist policymakers in reaching thoughtful, informed decisions about the feasibility and desirability of state-level NAEP.

Lee S. Shulman
President, National Academy of Education
and Charles E. Ducommun Professor of Education

Table of Contents

Transmittal Letter ix

Acknowledgments xi

The Panel xiii

Executive Summary xv

1 *Introduction* 1

2 *The Conduct of the 1990 Assessment* 7

3 *The Content Validity of the Assessment* 19

4 *Scoring and Interpretation* 31

5 *Reporting the Results to the Public and the Press* 43

6 *The Future of State NAEP* 59

Appendix: Synopses of Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment 69

References 103

Acronyms ..... Inside Back Cover

Detailed Table of Contents

Transmittal Letter ix
Acknowledgments xi
The Panel xii

Executive Summary xv

1 Introduction 1

NAEP as an Independent Monitor of Trends in Achievement 2
The Value of State NAEP 3
The Challenge to NAEP Today 4
Evaluation of the Trial State Assessment 4
An Overview of this Report 4

2 The Conduct of the 1990 Assessment 7

Overview 7
The Trial State Assessment Sample 8
Sampling of Schools 9
Sampling of Students 9
Participation 10
Weighting and Adjustments for Nonresponse 11
Excluded Students 11
Private School Students 12
Individual Education Plan (IEP) Students 12
Limited English Proficiency (LEP) Students 13
The Implementation of the 1990 TSA 14
Training and Monitoring of Local Administrators 14
Session Administration 14
Student Motivation 16
Feasibility of State NAEP 17
State NAEP's Impact on National NAEP 17
Perspective of the State Testing Directors 18

3 The Content Validity of the Assessment 19

Overview 19
The 1990 Mathematics Objectives and Framework 20
The Adequacy of the Consensus Process and the Framework 22
The Fit of the Items to the NAEP Framework 23
Implications for Future Assessments 26
The Influence of the NCTM Standards 26
The Need for Balance in NAEP Content and Assessment Methodologies 28
Directing the NAEP Development Effort 29

4 Scoring and Interpretation 31

Overview 31
The Use of Scale Scores, Anchor Items, and Anchor Levels 32
Achievement Levels 33
The Educational Standard-Setting Movement in the United States 33
What Are Standards? 34
Content Standards 34
Performance Standards 34
How NAGB Set Achievement Levels 35
The Need to Set Standards in a Relevant Context 36
How Useful Are the Achievement Levels for Reporting Educational Progress? 37
The National Educational Goals Panel's Decision to Report Competency Levels Instead of Achievement Levels 40
The Need for Further Research and Study on the Setting of Educational Standards and Achievement Levels 40
The Need to Integrate Setting Achievement Levels with Other Standard-Setting Activities in the Nation 41

5 Reporting the Results to the Public and to the Press 43

Overview 43
The Timetable for the Release 44
The Results Released in June 46
State-by-State Comparisons 47
Grade Level Equivalents for NAEP Scale Anchor Levels 50
What the Press Reported 51
State Comparisons 52
Press Coverage and the NCES Release Schedule 53
The Achievement Levels and the National Goals 53
How Useful Did the States Find the Reports of the Results? 56

6 The Future of State NAEP 59

Overview 59
Should State NAEP Be Continued and If So, in What Form? 60
Should NAEP Provide an Assessment of the Educational Achievement of All of the Nation's School-Age Children? 61
Should the Prohibition Against Reporting NAEP Scores Below the State Level Be Lifted? 62
Arguments in Favor of Lifting the Prohibition 63
Arguments Against Lifting the Prohibition 63
What Role Should NAEP Play in Education Reform? 65
NAEP Can Be a Monitor of the Status of Education at the National Level 65
NAEP Can Monitor the Results of State-Level Education Reforms 65
State NAEP Can Be Useful in Helping to Identify Strengths and Weaknesses in Education Systems 66
Summary 67

Appendix: Synopses of Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment 69

A Critique of Sampling in the 1990 Trial State Assessment 70
Eligibility/Exclusion Issues in the 1990 Trial State Assessment 73
Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment 76
The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis 80
The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns 84
A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report 87
General Issues in Reporting of the NAEP Trial State Assessment Results 91
The Case for District- and School-Level Results from NAEP 94
Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting Below the State Level 97
Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot be Made 101

References 103

Acronyms ..... Inside back cover

March 1992

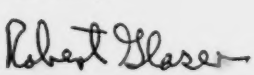
Emerson Elliott, Acting Commissioner
National Center for Education Statistics
U.S. Department of Education
555 New Jersey Avenue NW
Washington, D.C. 20001

Dear Commissioner Elliott:

On behalf of the National Academy of Education, we are pleased to transmit to you this Congressionally mandated report on the evaluation of the NAEP Trial State Assessment Project. Over the past two years, the Academy's Panel has studied technical and policy issues that have surrounded the development and implementation of the state trials. In April of 1991, it issued an interim report that offered recommendations, based on its preliminary research and deliberations, to help inform decisions then being made by Congress on the reauthorization of state NAEP. This document is the Panel's first mandated report. It presents the Panel's findings and observations on the 1990 mathematics trial state assessment program and surrounding events, and includes its recommendations regarding the continuation of the trials for consideration by the Congress, the states and the Executive Branch. The Panel then offers its perspective on the future role of NAEP in American education.

Throughout this first phase of its evaluation of the trials, the Panel has found it essential to review NAEP in a rapidly changing context of activities, proposals for national assessments and examinations, and myriad technical factors that can affect the validity and interpretation of indicators of educational change. In the four years since the report of the Alexander-James Study Group, which set the stage for NAEP administration and reporting at the state level, national attention to education reform has intensified. Numerous innovations that aim to affect the quality of American schools well into the next century are now being implemented at the national, state and local levels. Information from well-conceived assessments will therefore be indispensable to decisionmakers throughout our education system, as they evaluate and refine reform activities. Indeed, the state NAEP trials are occurring during a period when we can expect heightened interest in the potential uses of various kinds of tests and examinations as catalysts for reform. We hope that provision for the evaluation of the TSA is an indication that trials and systematic study of innovations in assessment will be an essential part of the changes in nationwide testing and in American education over the next years.

Sincerely,



Robert Glaser
Chairman

Robert Linn
Chairman

The National Academy of Education Panel on the
Evaluation of the NAEP Trial State Assessment Project

Acknowledgments

This mandated report presents the findings of the Panel of the National Academy of Education regarding the 1990 Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP).

The Panel is grateful to the National Center for Education Statistics (NCES) for its responsiveness in facilitating the Panel's independent review. Emerson Elliott, Acting Commissioner, was generous with his time and abundant knowledge whenever the chairmen and Panel sought information. Gene Owen, Chief, Operations and Information Branch, facilitated a productive and harmonious working relationship between NCES and the Panel. Other NCES staff who assisted in the production of the report include Steve Gorman and Gary Phillips, who were particularly helpful with statistical inquiries; and Ed Mooney and Tongsoo Song, who helped oversee the administration of the grant and fulfillment of its contractual details.

The contractors for the 1990 NAEP assessment, Educational Testing Service (ETS), Westat, and National Computer Systems (NCS), were of great help to both the Panel and our investigators. Among the numerous ETS staff who promptly and courteously responded to the Panel's requests for data and technical assistance were (in alphabetical order): Jules Goodison, Eugene Johnson, Steve Koffler, John Mazzeo, Ina Mullis, John Olson, Don Rock, and Alfred Rogers. At Westat, Nancy Caldwell and Keith Rust graciously answered questions and supplied vast amounts of materials and information, and Chris Schroeder played an important role in arranging observations of training sessions and classroom administrations of the first TSA. Lynn Zaback at NCS provided Panel staff with important data from the TSA.

The members of the National Assessment Governing Board (NAGB) staff deserve special thanks for their assistance, particularly Roy Truby, Dan Taylor, Mary Lyn Bourque, and Mary Crovo. At the Council of Chief State School Officers (CCSSO), Ramsay Selden and Rolf Blank aided the Panel in its research efforts. Kevin Crowe of CCSSO's Education Information Assessment Committee was especially helpful in providing information and inviting staff to relevant committee meetings.

The Panel sincerely appreciates the academic and government specialists who spoke at Panel meetings, providing information which ranged form historic perspectives to leading research on the TSA. The Panel thanks Albert Beaton, Chester Finn, Terry Hartle, Andrew Hartmen, Jack Jennings, Chancey Jones, Mary Lindquist, David Pearson, and Renee Slobaski.

Finally, the Panel expresses its gratitude to the distinguished researchers who carried out the commissioned investigations. Summaries of the commissioned reports are included in the appendix of this volume. The reports provided the Panel with in-depth knowledge of critical aspects of the assessment; the investigators played a vital and continuing role as advisors in their areas of expertise.

The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project

Robert Glaser, Chairman
Director, Learning Research and Development Center
and National Research Center on Student Learning
University of Pittsburgh

Robert Linn, Chairman
Co-director, National Center for Research
on Evaluation, Standards and Student Testing
University of Colorado at Boulder

Gordon Ambach
Executive Director
Council of Chief State School Officers

Isabel Beck
Professor of Education
Senior Scientist, LRDC
University of Pittsburgh

Lloyd Bond
Professor, Educational Research
Methodology
University of North Carolina at
Greensboro

Ann Brown
Professor of Education in Math, Science,
and Technology
Evelyn Lois Corey Fellow in Instructional
Science
University of California at Berkeley

Iris Carl
Houston Independent School District
President
National Council of Teachers of
Mathematics (NCTM)

Linda Darling-Hammond
Professor of Education
Columbia University

Robert M. Groves
Associate Director for Statistical Design
Methodology and Standards
U.S. Census Bureau

Lyle Jones
Professor of Psychology and Director,
L.L. Thurstone Psychometric Laboratory
University of North Carolina at
Chapel Hill

Edward Roeber
Director of Student Assessment Programs
Council of Chief State School Officers

Albert Shanker
President
American Federation of Teachers

David K. Cohen
John A. Hannah Distinguished
Professor of Education
Michigan State University

Ramon Cortines
Superintendent of Schools
San Francisco School District

Alonzo Crim
Benjamin E. Mays Professor of
Urban Educational Leadership
Georgia State University and
Professor of Education
Spellman College

Lorrie Shepard
Professor of Education
University of Colorado at Boulder

Marshall Smith
Dean, School of Education
Stanford University

William Winter
Former Governor
State of Mississippi

Project Staff

American Institutes for Research

George Bohrnstedt, Project Director
Phyllis DuBois
Elizabeth Hartka
Donald McLaughlin
Josephine Morrissey
Frances Stancavage

Learning Research and Development Center

Michelle von Koch

Executive Summary

Assessing Student Achievement in the States

The First Report of the National Academy of Education
Panel on the Evaluation of the NAEP Trial State Assessment Project:
1990 Trial State Assessment

March 1992

The summer of 1990 saw the twentieth anniversary of the release of the results of the first National Assessment of Educational Progress. Since 1970, more than 200 NAEP reports have been published. The first report of NAEP results, published in July 1970, was entitled, simply: "Science—National Results, National Assessment of Educational Progress Report, Number 1." The September 1990 report, which summarizes the results of the first two decades, has a more interpretive title: "America's Challenge: Accelerating Academic Achievement." The change in tone is hardly coincidental. Expectations for NAEP have risen sharply, especially in recent years.

"The Nation's Report Card," as NAEP has come to be known, more and more frequently is called upon to serve new roles. In 1988, Congress, the Administration, the Nation's governors, the chief state school officers and other education professionals determined that the time had come to see whether the nation's report card might also become a report card for the states. In the spring of that year, Congress enacted Public Law 100-297, authorizing a NAEP Trial State Assessment (TSA) program to determine whether state assessments following the NAEP format could produce reliable and useful estimates of educational progress and information about the several state education programs.

The Context for the 1990 Trial State Assessment

The authorization from P.L. 100-297 for expanded data collection on educational attainment using NAEP came at a time of intense discussion and debate about education reform and the role assessment should play in promoting reform. Then, following the September 1989 Education Summit at Charlottesville, Virginia, the Administration, working with the Nation's governors, formed the National Education Goals Panel (NEGP). The panel agreed to issue their own "report cards" annually so that the nation and the states could measure progress toward achieving the six National Education Goals that had been developed at Charlottesville, and which were adopted in February 1990 by the President and the governors. Goal Three was of particular importance for NAEP because it states that by the year 2000 American students should leave the 4th, 8th, and 12th grades having demonstrated competency in English, mathematics, science, history, and geography. Shortly thereafter, several groups called for national examinations to provide direct evidence to parents about what their children know and can do, thereby serving as a stimulus for changes in

education. The push for national examinations became linked with a movement to reform educational assessment through tests that are more closely integrated with curriculum and instruction, as well as with concepts of assessment that would reflect higher order thinking and problem solving. Several groups, including NAEP's National Assessment Governing Board (NAGB), also began arguing for the establishment of national education standards and testing. In June 1991 Congress passed the legislation that created the National Council for Education Standards and Testing (NCEST), a body charged with providing the President, Congress, and the governors with advice on the desirability and feasibility of national education standards and testing. Advisory groups to the NEGP distinguished between examinations intended to reform educational practice and monitoring systems such as NAEP, while other groups have argued that NAEP should be used to serve both purposes. Each of these developments has raised the stakes for NAEP and the TSA and transformed expectations about the role that state NAEP might play in education change and reform.

The Congressional Requirement for an Independent Evaluation of the TSA

In authorizing state NAEP, Congress called for an independent evaluation of "the feasibility and validity of [state] assessments and the fairness and accuracy of the data they produce." Congress viewed the evaluation of a set of trials in the states as prerequisite to establishing a permanent NAEP program at the state level.

In October 1989, the National Center for Education Statistics (NCES) commissioned the National Academy of Education to conduct the evaluation. The Academy appointed an independent Panel, chaired by Professors Robert Glaser and Robert Linn and composed of a broad spectrum of technical and policy experts in education, to undertake the work. It also arranged for technical and staff support from the American Institutes for Research with Dr. George Bohrnstedt serving as Project Director. The Panel, as it is referred to throughout this document, commissioned a set of papers and studies that provide much of the background and data for this, the first of three mandated reports on the state NAEP program.

In February 1990, the first trial of the state NAEP, an assessment of mathematics achievement, was administered in more than 3,500 schools to some 100,000 of the nation's eighth graders. In total, 37 states, the District of Columbia, and two territories participated, indicating wide interest in state NAEP. The results of the 1990 trial were released on June 6, 1991. The second trial is scheduled for 1992, with the data collection to be expanded to include fourth grade reading, as well as fourth and eighth grade mathematics. The continuation of state NAEP in 1994, however, is contingent on Congressional reauthorization.

The key questions that have guided the Panel's work are those that Congress, in 1988, anticipated would be crucial to evaluating the TSA: (1) How well was the assessment implemented from a technical perspective? (2) How valid and accurate is the assessment? Has it yielded valid and reliable data at the state level? (3) How useful are the results and reports generated from the assessment?

The Panel's Findings and Recommendations

Through its deliberations and its research, the Panel has developed a set of recommendations to Congress, the states, and the Executive Branch. These recommendations are grounded in studies of the 1990 trial and address major issues that can help ensure the effectiveness of state NAEP in the future. The Panel offers its recommendations in hopes that they will inform Congressional deliberation on specific aspects of the reauthorization of NAEP that will have import for the future not only of NAEP and the state program, but for assessments more broadly.

Should State NAEP be Continued?

The immediate question before Congress is whether or not state NAEP should be continued. Although the Panel found that the 1990 TSA was carried out successfully and with a high degree of validity, this was but a single trial of a program that would necessarily become a much larger undertaking if state NAEP were fully implemented.

On the basis of its findings, the Panel believes that Congress should approve the continuation of state NAEP, but before legislating a permanent state NAEP, should authorize additional trials. Because only two subjects at grade four and one subject at grade eight will have been assessed at the conclusion of the 1992 TSA, and because expansion of the trials to more subjects at all grade levels would provide the data needed to evaluate the feasibility, impact, and cost of a more fully implemented state NAEP, the Panel recommends the continuation of the trial program in 1994, rather than the full establishment of a state NAEP program. In particular, the Panel recommends trials in 1994 in the 4th, 8th, and 12th grades in mathematics, reading, and one additional subject, such as science.

Should the Prohibition Against Reporting NAEP Scores Below the State Level be Lifted?

Perhaps the second most pressing issue Congress must confront is whether it should lift the prohibition against reporting NAEP results below the state level. The Panel has considered this issue very carefully, knowing that Congress has received recommendations to remove the ban.

After weighing the reasons for and against lifting the prohibition, the Panel concludes that NAEP has been and should continue to be a *global* monitor of educational achievement. Its historic role as an independent auditor that neither is unduly influenced by particular instructional or testing practices nor shapes day-to-day classroom activities would be lost if TSA results were to be reported at the district, school, or student levels. In addition to compromising NAEP's value as an indicator, lifting the prohibition could detract from local innovations that attempt to integrate assessments with

curricula. Therefore, the Panel concludes that the prohibition against reporting NAEP data below the state level should be continued. [1]

In addition to these two questions, the Panel addressed issues regarding the content validity, administration, implementation, exclusions, and the reporting of the TSA.

The Content Validity of the 1990 Assessment

In examining how the content of the 1990 mathematics assessment was determined, *the Panel concludes that the 1990 NAEP mathematics framework and the assessment design principles on which it was built were essentially sound.* The designers of the 1990 framework appear to have captured the spirit of change in mathematics education during the late 1980s while also preserving traditional elements; they thus created a reasonable compromise between the visions of educational reformers, reflected in the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics,* and the realities of educational practice for that time period.

It is important that NAEP's content be comprehensive, reflecting the most forward-looking, modern approaches, at the same time that it reflects the best of current practice. The Panel applauds the increased testing of higher order thinking skills and challenging concepts in the 1990 TSA and recommends that all NAEP content frameworks evolve in this direction. Similar moves should be made to incorporate cutting-edge assessment techniques, at the same time retaining the best of more traditional methodologies. For future mathematics assessments, the Panel recommends including items that assess *all* the themes identified in the NCTM Standards including innovative items that require extended written responses.

A reasonable effort was made to include a wide array of relevant stakeholders in the consensus process for the development of the framework, but the Panel is concerned about the lack of continuity between the groups responsible for developing the framework and those who wrote items and determined how they should be scored.

The development of assessments in the various NAEP content areas is a multiphase process, involving numerous technical and policy groups. As a result, the process is subject to fragmentation. The Panel recommends that a mechanism be devised to assure continuity throughout, beginning with the development of the framework and continuing through the various stages of item specifications, item writing, item scoring, and reporting of the results. This mechanism might be a set of oversight committees, one for each subject matter area, composed of both NAGB members and experts from the subject area communities. These committees would help to ensure effective and on-going communication with those most knowledgeable in the NAEP content areas.

The items that appeared on the 1990 mathematics assessment reflected the NAEP content framework well. As a result, *the Panel concludes that the assessment provides a valid reflection of the NAEP content framework.* The fit of the items to the

[1] Ambach dissents from this position; he is on record elsewhere as recommending lifting the prohibition at the school district level where the size of the enrollment enables sampling as used at the state level.

framework implied by the NCTM standards was not as good, but this is not surprising. The NCTM framework is new and was designed to raise standards and encourage practices not commonly observed in mathematics instruction. Overall, the 1990 assessment took an important step towards the development of a *comprehensive* mathematics assessment. However, *the Panel believes that assessment frameworks must be viewed as evolutionary entities rather than as structures that are fixed for long periods of time*. The framework developed for 1990 must therefore be allowed to evolve to reflect changing conceptions of mathematics education in the nation.

The Panel's study of the development of the 1990 mathematics assessment makes it clear that planning must begin well in advance for future assessments so that adequate time can be allowed to develop exemplary assessment tasks and to find ways to measure constructs not well covered by today's assessment methodologies. Participants in the mathematics development process felt strongly that they did not have sufficient time to complete their tasks in a thoughtful manner. More recent efforts at developing content have continued to suffer from overly tight timelines. The Council of Chief State School Officers (CCSSO) had less than four months to create the 1992 reading framework and only seven months to create the 1994 science framework. One-year-at-a-time reauthorizations and extensions just prior to the scheduled field tests are not adequate. Congress should authorize NAEP assessments several years prior to their scheduled administration in order to resolve this ongoing problem.

Implementation, Administration and Exclusions

Implementation. The sampling design for the 1990 TSA met the high standards expected of a major national education survey. It was executed with care and diligence by the NAEP contractor, and the resulting samples were of sufficient size to warrant inferences about the populations sampled. Participation rates of schools and students selected for the sample were also generally good. The overall participation rate for the schools originally selected was about 94 percent; the average participation rates for sampled students ranged from 92 percent to 96 percent across the states, with two exceptions.

Administration. The assessment appears to have been carefully and consistently administered for nearly all students who participated. The performance of students in 1,918 randomly-drawn sessions that were monitored by the contractor did not differ significantly from the performance of students in 1,997 sessions that were not monitored.

Monitoring the administration of the 1990 mathematics assessment provided valuable data and helped to assure uniform testing conditions. As the TSA expands into more states, and different administrators are trained to carry out the testing, the Panel recommends that random monitoring of schools be continued. Uniform procedures for administration are essential for helping assure comparable results across states.

Exclusions. P.L. 100-297 specifies that all public school students who are selected for the TSA and are capable of participating should be assessed, but it allows for the exclusion of three groups of students: students who are enrolled in private schools, and students who have limited English proficiency (LEP) or individualized education plans (IEP) and are judged incapable of participating meaningfully in the assessment.

Private school students typically made up the largest group of excluded students in each state. Nationally, about 12 percent of the eighth grade students are enrolled in private schools, but state percentages vary from less than 5 percent to about 20 percent. Eighth grade private school students, on the average, score somewhat higher on NAEP than those enrolled in public schools. Analyses indicate that the relative standing for some of the states would have been changed noticeably if private school students had been included.

State NAEP data would better reflect educational achievement and make state results more readily comparable if results for all students, private as well as public, were included. Ultimately, because of the significant variation from state to state in private school enrollment, the exclusion of private school students from the TSA will diminish its utility. The Panel therefore recommends that future authorizations for state NAEP include adequate resources to sample private school students, thereby increasing the comparability of results from one state to another.

The exclusion criteria for IEP and LEP students were part of the sampling design of the TSA; the actual exclusion decisions were made by local school personnel applying the prescribed criteria. Overall, about four percent of the sampled students were excluded because they had individualized education plans and were evaluated as not able to participate, but the exclusions ranged from two percent to eight percent across the states.

Significant variations in the state-by-state exclusions of IEP students were observed in the 1990 TSA that cannot be easily explained. Importantly, the Panel's research shows that differential exclusion rates affect the rankings of the states. Therefore, the Panel recommends that NCES conduct a study designed to evaluate the rationales used by educators for the exclusion of IEP students on the basis of their ability to participate meaningfully in the assessment. This study would result in a better understanding of the differential use of exclusion criteria across states, thereby providing information that would allow states to compare themselves more accurately on NAEP assessments.

In theory, NAEP should include a representative sample of all of our 4th, 8th, and 12th graders; the principle of inclusion is an important one. With the increase in immigration to our nation, substantial numbers of school-age children do not read English well enough to allow them to be assessed by NAEP. Among sampled eighth graders, the overall rate of exclusion due to limitations of English proficiency was one percent, but varied from near zero to five percent across the states. This variation seems reasonable given the differing demographic patterns in the states. However, because of the increasing number of students whose English speaking abilities are limited, *the Panel believes that serious consideration should be given to the development of non-English versions of NAEP for implementation on a trial basis.* To encourage attention to this issue, the Panel is commissioning a study on the feasibility, cost, and advisability of non-English versions of the mathematics assessment as part of its continuing evaluation of the TSA. The Panel will report its findings, with its evaluation of the 1992 trials, to the acting commissioner of the NCES with the purpose of providing them to Congress.

The Panel also raised the issue of the exclusion of school dropouts. At present, NAEP is an assessment of what children *who are enrolled in school* at the 4th, 8th, and 12th grade levels know. It is well-known that a significant proportion of our children drop out before completing the 12th grade. We also know that children who drop out have lower standardized tests scores and lower grade point averages while they are in school. Moreover, because the dropout rate has changed and continues to change across time, the composition of the group of students enrolled in school and therefore assessed by NAEP continues to change. *These continuous changes in the composition of the school-age population call into question the validity of current comparisons of the educational achievements of 12th graders across time.*

While such validity issues are of paramount importance when making comparisons of 12th graders across time, perhaps of more importance is the paucity of information about the knowledge and skills of that proportion of the population of children of 12th-grade age who have dropped out of school and are not currently assessed—a group of young people at high risk for crime, delinquency, drug use, and welfare status. Although it is assumed that their employability is low because they lack the requisite knowledge and skills, reliable and valid data with which to compare their educational achievement against that of young persons of the same age who are enrolled in the 12th grade do not exist.

Providing a reliable picture of what all school-aged children know has been the stated aim of NAEP since its inception. Valid comparisons of assessment results across time require that samples of 17-year-olds who are no longer enrolled in school be assessed routinely. The Panel recommends that beginning with the 1994 assessment, a national sample of out-of-school 17-year-olds be assessed. This approach would be less expensive than testing state samples, but should be supplemented by giving each state participating in the TSA the option of paying for the assessment of a similar sample of its out-of-school 17-year-olds.

Score Reporting and Educational Standards

The results from the 1990 trial were released in two different formats—the traditional NAEP scale scores, with highlighted anchor levels, and NAGB's new achievement levels, which are intended to define what students at each grade level should be able to do.

NAEP Scale Score Reporting. The NAEP scale score results using the traditional anchor points of 200, 250, 300, and 350, along with average NAEP scale scores, were released to the public on June 6, 1991. NCES and its contractor, the Educational Testing Service (ETS), did an excellent job of presenting the results of the 1990 TSA to the public. The assessment yielded an enormous wealth of data that challenges the reader in terms of both its sheer volume and technical complexity. The decision to release multiple reports, each at a different level of summarization, helped enormously to keep the data manageable. NCES and ETS are also to be commended for their work in presenting the state level data in accurate and fair ways, while at the same time developing reporting formats that discouraged simplistic conclusions about state comparisons.

The central message of the June 1991 data release—that mathematics performance in the United States is inadequate and needs to be improved—was, in general, reported accurately in the press coverage of the results. A high proportion of the published articles, however, discussed the state rankings without providing the statistical information necessary to help the reader interpret results for states with overlapping levels of performance. In addition, significant coverage of some of the most potentially useful findings was lost in the reporting; for example, results that examined the relationship of student performances to differences in mathematics curriculum and teacher preparation.

The Panel recommends that NCES continue to provide the media with specific examples of interpretations that are (or are not) supported by the results as well as continue efforts to construct new tables and graphs that effectively display the data. The Panel also recommends that newly developed reporting devices be presented for comment to the press or media specialists in a series of focus groups, prior to the release of future reports. These efforts may enable NCES to determine the types of data and displays that best satisfy the reporting needs of the media while preserving the accuracy of the data.

The Panel also suggests that NCES consider spreading the release of the data over several months. After the first report on the state-by-state results, individual reports could be released concentrating on the relationship of outcomes with such variables as course-taking patterns, instructional time devoted to particular subjects, teacher experience and preparation, and school characteristics. It would be important, however, to insure that the initial report be sufficiently inclusive to provide a fair context for the comparisons between the states.

The June release of results provoked discussions in state departments of education, state boards of education, and state-level educational organizations. State mathematics specialists and testing directors interviewed in late summer 1991 reported that the results were providing support for reforms that were contemplated or already underway. They also indicated that the impact of the state assessments could be expected to increase once successive rounds of testing had occurred and the states could review their performances over time.

Achievement Level Reports. In late 1989, NAGB began a discussion about the merits of setting achievement levels for the nation's 4th, 8th, and 12th graders. It was argued that instead of reporting what our students can do (i.e., through the use of the traditional anchor levels), it was more important to report achievement in the context of what the nation's 4th, 8th, and 12th graders *should* be able to do. In May 1990, NAGB passed a set of policy guidelines for setting achievement levels for the nation, citing its statutory responsibility to identify appropriate achievement goals for the nation's school children. Three achievement levels were established: basic, proficient, and advanced.

The achievement-levels-setting effort was closely scrutinized by a number of parties including, but not limited to, NAGB's own evaluation team, the NCES Technical Review Panel, CCSSO, and this Panel. The adequacy of NAGB's definitions of achievement levels has been a source of considerable controversy. The NAGB evaluation team concluded that (a) the achievement levels should not be used as a "baseline or benchmark" against which to compare future achievement level results and (b) the procedures used to set the levels should not be treated as a model for

future NAEP assessments. Contrary to the report of its evaluation team, NAGB adopted the achievement levels. At the request of Congress, this matter is being reviewed by the General Accounting Office. It is the Panel's understanding, however, that the achievement levels data released on September 30, 1991, *will not* be used by NAGB as a baseline against which to compare future achievement level results *unless* additional evidence is garnered to warrant such a use. *The Panel strongly endorses this position.*

It is NAGB's intention to use the achievement levels to establish a new trend line beginning in 1992. The Panel believes that the utility of reporting what students should know compared to what they do know has yet to be demonstrated. The Panel is in the process of gathering information from policymakers and educators at the national, state, and local levels to inform a recommendation about the usefulness of the achievement levels as a reporting device. The Panel will report the results in its evaluation of the 1992 TSA. In the meantime, the Panel strongly suggests that NCES report the results from the 1992 assessment using the NAEP scale score and its anchor levels, whether or not the 1992 achievement levels are used.

As part of its first-ever report card to the states, the NEGP reported the NAEP achievement level results on September 30, 1991, in terms of "competency levels." "Competent" was defined by NEGP as the combination of NAGB's proficient and advanced achievement levels and "not competent" as the combination of the basic and below basic levels. The Panel believes that little was gained by collapsing the data in this way; indeed, it may have increased ambiguity about the meaning of the data. Because of the potential for confusion, the Panel suggests that NCES, NAGB, and NEGP agree on a single format for releasing the achievement levels data in future years.

Much of the current reform movement in mathematics is based on the assumption that there are multiple aspects to mathematics achievement. As a result, mathematics educators report finding the NAEP subscores more useful than the global NAEP score. Because NAGB has recommended that, beginning in 1992, the achievement levels are to be the primary vehicle for reporting NAEP results, the subscores and their value to educators are in danger of being lost—a serious loss in the Panel's view. The Panel strongly suggests that the results of the 1992 assessment be reported by subscore as well as by the total NAEP scale score.

Finally, the Panel is concerned that the plans for the 1992 achievement-levels-setting process do not call for an external evaluation. Given the problems that occurred with the 1990 assessment and the great importance of the achievement-levels-setting process, the procedures for setting achievement levels should be evaluated by a distinguished team of external evaluators as well as by an internal team.

In June 1991, the momentum for the development of national standards was given a strong boost when Congress, working with the governors and the Administration, created NCEST to examine the desirability and feasibility of developing national standards. In its December 31, 1991, report to Congress, the Council endorsed the setting of national content and performance standards. NCEST also recommended that Congress establish a new council on national standards and assessment to coordinate the development of new education standards and then to certify that they are "world class." Assuming that Congress enacts NCEST's recommendation, it is

very important that NAGB, in the development of the achievement levels, work closely with NCEST's successor.

The use of achievement levels in reporting NAEP results can be of value both to the nation and to the states participating in the TSA, by providing succinct, accessible information for public discussion of education and reform. However, it is essential that NAGB's future efforts at setting achievement levels be closely coordinated with the reports and recommendations of the council proposed to succeed NCEST. This will be essential to ensure that a coordinated set of educational standards informs policy and discussion. The Panel further recommends that when setting achievement levels for future assessments, NAGB conduct studies to validate the descriptors used to illustrate the achievement levels—especially those implying outcomes of future performances, such as success in postsecondary education and employment.

The Role of State NAEP in Education Reform

In this time of anticipated educational change in our nation, NAEP and the trial state NAEP can become indispensable sources of information. Because it generates trend data over a regularly spaced interval of years, NAEP can be useful to the states and to the nation in helping to gauge the extent of educational progress. State NAEP can pinpoint both inequities and strengths in our various education systems. In the context of our current ambitions for education reform, including proposals for new examination systems at the local, state, and national levels, *NAEP should be preserved as the nation's key independent indicator of how students are performing academically. NAEP has played and should play the same role for education that economic indicators have played in signaling change in the status of the economy.* NAEP can also serve as a model for the kinds of content, forms of competence, and levels of achievement that reflect national goals. The state trials and the study of them provide an opportunity for building improved indicators of our children's educational attainment—indicators that can mirror and inform ongoing and future reforms in our education systems.

1 Introduction

This first report of the National Academy of Education's panel on the evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment (TSA) comes at a time of precedent-setting discussions on the role of assessment in education reform. New policies in education are moving in unanticipated, even previously opposed, directions. The decision to have the National Center for Education Statistics (NCES) collect state-by-state achievement data itself represented the reversal of longstanding policy; it now appears that there is widespread agreement on the value of accurate and valid state-level measures of educational achievement. Public Law 100-297 took the first step towards this goal when it provided for the voluntary participation of states in a NAEP state-level assessment on a trial basis in 1990 and 1992.

The state trials began amid a proliferation of activity focussing on national goals and standards, curriculum frameworks, and nationwide assessment of individual student achievement. At the September 1989 Education Summit—one year after P.L. 100-297 passed—the Nation's governors and the President decided to set national goals for education. The goals were agreed upon in February 1990, and in July 1990 the National Education Goals Panel (NEGP) was formed to measure progress towards these six goals by issuing annual "report cards" for the nation and the states. An authorization for voluntary national examinations had been enacted by Congress in 1988, but had never been implemented. By 1991, several groups, including the President's Education Policy Advisory Committee and the Council of Chief State School Officers, called for a national examination system. The Administration's *America 2000* education strategy featured a proposal for the American Achievement Tests.[1] In some of these proposals, NAEP was suggested for use as an interim examination.

During this same time period, the National Assessment Governing Board (NAGB) developed achievement standards for the NAEP mathematics assessment, and the National Education Goals Panel (NEGP) called for the establishment of standards of student performance for the proposed national examination system. More recently, Congress created the National Council on Education Standards and Testing (NCEST), responsible for making recommendations on the feasibility and desirability of national standards and assessments.

Thus, the first TSA in 1990 emerged as only one of several activities focussed on student achievement. Implicitly and explicitly, these other activities raise questions about the role that state NAEP can and should have in the overall design of nationwide student assessment and in education reform.

The assessment of student achievement can play two major roles in national efforts toward education reform. First, a nationwide examination measuring individual student achievement could become part of an integrated curriculum-assessment system—a system characterized by a continuity of goals, instruction, and standards widely endorsed by the nation's schools. Second, indicators like NAEP could produce separate national and state-wide estimates and benchmarks to inform parents, school professionals, policymakers and the public about the performance of students over time.

[1] U.S. Department of Education, *America 2000: An Education Strategy* (Washington, D.C.: 1991).

In the Panel's view, these two roles of assessment serve distinct purposes—purposes that are frequently confused. On one hand, a proposed national examination system is intended to certify student achievement on a body of knowledge and skills closely related to the school curriculum. Some believe such a system could set forth a set of clearly-defined goals towards which teachers can teach and students can study. On the other hand, NAEP is designed to be a more neutral public indicator of achievement that can span both what the schools are currently teaching and what they should attempt to teach. By taking a stance further removed from day-by-day practice, while still assessing similar standards and forms of competency, NAEP can serve as an important, impartial monitor of the quality and accomplishments of many of our states' education systems, particularly as they change.

At this time of heightened aspirations for testing, it is well to remind ourselves that major systemic changes do not occur in a miraculous instant or even in a prescribed time span. Productive and lasting changes are achieved incrementally through a process of development, trial, evaluation, and redesign. We must recognize the inevitability of this iterative cycle even while we attempt to hasten it by bringing to bear appropriate resources and a sense of national urgency. Congress recognized this important fact in its legislation on state NAEP. Thus it authorized a *trial* assessment that would be evaluated over a period of years while its effects were being ascertained, and its quality and utility, analyzed and improved.

NAEP as an Independent Monitor of Trends in Achievement

For more than 20 years, NAEP has been a continuous indicator of the status of the nation's education system, providing invaluable trend data on the progress of American elementary and secondary students. As the Alexander-James report stated:

> [NAEP] has become the primary source of information on what American students, nationwide, know and can do, covering a wide range of school subjects....Since by design that information is comparable in many respects from one assessment to the next, we can now compare the children of today with those of earlier years and learn how well our educational system is working. No other measure of educational achievement has NAEP's size, scope, and perspective over time. [2]

Further, unlike results from college admissions tests such as the Scholastic Aptitude Test (SAT) that are often used inappropriately as indicators of shifts in nationwide achievement, NAEP scores represent the achievement of nearly all students at three grade levels, not just that subset of high school seniors who are college-bound.

Because of NAEP we know that, despite declines in college admissions test scores, the average achievement levels in the core disciplines of reading and mathematics have been quite stable from 1969 to the present. We cannot take pride, however, in this stability. The levels that have held constant over the past 20 years are below the levels required for students' successful participation in the workforce of the future and

[2] National Academy of Education, L. Alexander and H. T. James. *The Nation's Report Card: Improving the Assessment of Student Achievement* (Washington, D.C.: 1987), 11.

for the well-being of the nation. Education Secretary Lamar Alexander emphasized at the release of the NAEP trend data on September 30, 1991, that, because of changes in technology and the economy, "What we did in 1970 is not nearly good enough in 1990." [3] Even more distressing is the performance of 17-year-olds in science, where NAEP has indicated a significant *decrease* in achievement over the past 20 years. Across the disciplines, NAEP results have also shown us that students are failing to master higher order skills even when they are succeeding in relation to more limited procedural skills; taken together, these findings have helped to focus attention on the kinds of changes needed in school curricula.

NAEP has been the source of encouraging information as well. It has provided valuable insights into positive changes in achievement by race, ethnicity, and sex. Through NAEP, policymakers learned in the 1970s and 1980s that minority students have significantly narrowed the gap between their academic achievement and that of whites, although that gap remains unacceptably large.

The Value of State NAEP

Just as national NAEP has informed the nation about the educational achievement of U.S. students over time, state NAEP can begin to inform each state about the achievement trends of its own students. States that choose to participate in state NAEP will be able to monitor the relative rate of change in student performance in several key subject areas because of the across time measurements state NAEP will provide. Although the state-level trend data will be of great value to those states looking for indications of the impact of major, statewide, systemic education reforms, it is also of great importance that the data be interpreted carefully since factors such as changes in a state's demographics, poverty level, or health care available for school-age children might also account for changes in performance on NAEP.

Because state NAEP results are reported for subpopulations, states can compare the performances of various racial-ethnic and socioeconomic groups, or they can compare students scoring at different levels. For example, because education reforms may affect different sectors of the population in different ways, states may wish to monitor longitudinal changes in performance for students who score in the top 10 percent and those who score in the bottom 10 percent as well as tracking trends in average performance. NAEP can also be of use to states that wish to compare the performance of subgroups of students with those same subgroups in other states. Thus California may be interested in comparing the performance of its Hispanic students with those in Texas, or Ohio may want to compare the performance of its disadvantaged urban students with those in Pennsylvania. As results from the 1990 TSA have already shown, the rankings of the states are somewhat different if they are based on the top 10 percent of students in each state rather than the average performance, or if the scores for different racial-ethnic groups are compared. State NAEP data can also serve to raise important questions about the relationship between achievement and such contextual factors as students' television watching habits or their teachers' training. Although such comparisons and observed relationships can be important to the states in suggesting areas for change in education practices, great care must be taken not to draw causal inferences when making such comparisons,

[3] Kenneth J. Cooper, "Education Goals Called Far from Realization: Report Will Serve as Baseline for Progress," *Washington Post* (October 1, 1991): A-4.

because variations in achievement both within states and across states can be due to a variety of factors in addition to schooling.

Finally, state NAEP enables states to compare the overall performance of their students to that of the nation as well as to comparable groups of students in other states. Together, national and state NAEP can offer valuable long-term measurements of student performance for both the nation and the participating states.

The Challenge to NAEP Today

With the current high level of expectations for assessment, NAEP is being called upon to play a larger role than ever before. For example, there are calls for NAEP to develop information below the state level to the district, school site, and even individual levels. In addition, there is significant pressure to accelerate the replacement of the current formats in the NAEP assessment with more innovative items to better reflect advances in cognitive science and measurement as well as the complex demands of real-life tasks. At the same time, some reformers have urged that NAEP incorporate new concepts of literacy—in mathematics, science, language arts and other core disciplines—thereby encouraging changes in curriculum and instruction rather than merely reflecting current teaching.

These pressures for continued reform are evidence of the urgent concern for improvement in the nation's educational achievement. They also present important challenges to NAEP's historic place in American education. Because the costs and benefits of the proposed reforms require careful analysis, the Panel recognizes that its evaluation of the first trials of state-level NAEP must be broad enough to encompass these concerns.

Evaluation of the Trial State Assessment

In authorizing state NAEP, Congress provided for an independent evaluation of "the feasibility and validity of [state] assessments and fairness and accuracy of the data they produce." The evaluation of a set of trials in the states would be prerequisite to establishing a permanent NAEP program at the state level. In October of 1989, the NCES commissioned the National Academy of Education to conduct this evaluation. The Academy appointed an independent Panel comprising a broad spectrum of technical and policy experts in education to undertake the work.

An Overview of this Report

To evaluate the 1990 trial, the Panel commissioned a set of papers and studies that provided much of the background and data for this, the first of three mandated reports that it will issue on the state NAEP program. The Panel has focussed on the

questions posed by Congress in 1988 as well as on newly emerging views of NAEP's role in education reform. Each of the five chapters that follow is designed to answer a set of questions. Following the Panel's report, synopses of the supporting reports and studies appear in the appendix.

Chapter 2 addresses the administration of the 1990 TSA. In it the Panel examines the adequacy of the design for state samples, asking if these samples accurately represented the eighth graders in the participating states or if the results of the TSA were in any way biased by inaccurate sampling or exclusion of important groups of students. The main question is: Are the results of the 1990 trial close to those that would have been obtained if every eighth grade student in participating states had been tested? The Panel also examines how well the assessment was implemented. Were the same procedures followed at each school? How many of the selected students actually participated? Were appropriate adjustments applied where some schools or students failed to participate?

The focus of Chapter 3 is the validity of the assessment's content. In it, the Panel asks: What aspects of mathematics did the 1990 assessment measure? Did the content framework that guided the development of the assessment accommodate an appropriate range of knowledge and skills? Was the consensus on the framework based on the broad-based input from mathematics educators that Congress prescribed? In this chapter the Panel also examines the extent to which the items in the assessment reflect the NAEP content framework and the more recent *Standards* of the National Council of Teachers of Mathematics (NCTM). Critically, did the assessment content provide the balance between current classroom practice and professional aspirations for mathematical literacy that is essential to preserve NAEP's role as an indicator?

The Panel examines the scoring and interpretation of the 1990 mathematics assessment in Chapter 4. The lead question is: How were student responses scored and what do the scores mean? The Panel then probes current scoring and reporting practices in relation to the appropriateness of NAGB's achievement levels and the relationship of NAEP scores to standards for mathematical performance. A central concern is whether the information provided by the achievement levels differs from that already available in the proficiency levels NCES has used for the past decades to report NAEP results. Are the achievement levels likely to prove more useful in reporting NAEP data? Are the achievement levels technically sound? Should state NAEP continue to be reported using these levels in future assessments?

In Chapter 5 the Panel examines the formats used for reporting TSA results to the public, policymakers and the press. How did NCES prepare the states for the results of the TSA? Were the results described in a thorough and comprehensible manner—one that promoted accurate reporting and meaningful interpretations of the results by the press? The Panel also asks whether the press, in reporting the results, used the full scope of information available. What were the main messages conveyed by the press when the results were first reported in June 1991 and when the results were again released—using the new achievement levels—in September? Finally, how useful have the states found the results? Has there been any significant effect on education policy during the first months following the release?

The sixth and final chapter contains the Panel's assessment of the future of state NAEP. Should the TSA be continued? If state NAEP is continued, what steps might be taken to improve it? Should state NAEP be expanded to include populations of young persons

not currently assessed? And finally, of paramount importance, given current calls for a national examination system and the emphasis on the role of assessment in education reform, what role, if any, should NAEP play in an expanded assessment arena?

2 The Conduct of the 1990 Assessment

Overview

Because NAEP data are used to make generalizations about how U.S. students as a whole are achieving, the adequacy of the sample of students tested is a critical issue. In carrying out its evaluation of the administration of the 1990 Trial State Assessment (TSA), the Panel therefore conducted a careful review of the adequacy of the sampling design and execution, including an examination of the effects of participation rates and intentional exclusions.

In addition, the Panel reviewed documents and conducted analyses to evaluate the quality of the assessment administration, recognizing that reasonable and uniform assessment conditions for all students are also important prerequisites for accurate data and defensible inferences. Unlike national NAEP, where contractor staff administer the assessment, the administration of the TSA was carried out by local school personnel. These new conditions of administration made it imperative to determine whether there were any important differences between states in the way that the assessment was administered.

Finally, the Panel reviewed participation data for the past several NAEP cycles to evaluate the impact of the TSA on national NAEP, and it surveyed state testing directors to determine the state's perspectives on the benefits and costs of administering the TSA.

Based on its examinations of the sampling design, documentation and other evidence concerning the administration, and the data generated from the assessment, the Panel concludes:

- *The sampling design used for the 1990 TSA met the high standards expected of a major national education survey. It was executed with care and diligence by the NAEP contractor, and the resulting samples were of sufficient size to draw reasonable inferences for the populations sampled.*

- *Participation rates of schools and students selected for the sample were generally good. The overall participation rate for the schools originally selected was about 94 percent; the average participation rates for sampled students ranged from 92 percent to 96 percent across the states, with two exceptions (noted below).*

- *Across the states, two to eight percent of the sampled students were excluded because of individualized education plans (IEP). These variations in state-by-state exclusions, which are not easily explained, could affect the rankings of the states.*

- *Overall, only one percent of sampled students were excluded because of limited English proficiency (LEP), but the rate of LEP exclusion varied from near zero to five percent depending upon the state. This variation seems reasonable given the differing demographics in the states.*

- *The assessment appears to have been carefully and consistently administered for nearly all students who participated. The performance of students in 1,918 randomly drawn sessions that were monitored by the contractor did not differ significantly from the performance of students in 1,997 sessions that were not monitored.*

- *There is no indication at this time that state NAEP assessments will have a negative impact on the national NAEP.*

- *Nearly three fourths of the state testing directors expected the results of the TSA to be worth the costs to their states. Nevertheless, many would have preferred to have had more opportunity for input on TSA policy decisions.*

The Trial State Assessment Sample

The goal of the Trial State Assessment (TSA) is to produce accurate, statistically defensible statements about the mathematics performance of eighth graders in each participating state, as well as comparable statements for selected subpopulations of the eighth graders within each state (for example, rural, urban advantaged, and urban disadvantaged eighth graders). In order to do this, one requirement is that the TSA be based on state probability samples that are of sufficient size to provide accurate estimates for each of these populations.

The quality of TSA findings depends not only on how well the sample is designed, but also on *how well the design is executed*. Efforts must be made to ensure high participation rates among selected schools and students, and valid estimation procedures must be employed to weight the data properly and to account for schools and students who were selected but failed to participate.

In addition, one must consider whether the validity of the assessment is seriously affected by design decisions that exclude certain categories of schools or students such as those who are hard to locate, difficult to test, or sparsely represented in the population. For the 1990 TSA three groups were excluded: private school students (who were included in the 1990 sample for national NAEP) and students with individualized education plans (IEP) or limited English proficiency (LEP), if they were considered unable to participate meaningfully in the assessment. (Individual education plans are prescribed by law for students with disabilities who are provided with special education services.)

Every exclusion limits the scope of generalization that can be derived from the findings. Thus, for example, the 1990 TSA results reflect only the performance of

public school eighth graders not so handicapped as to preclude meaningful testing on the NAEP instrument. Generalizations to *all* eighth graders in a given state are affected by the numbers of excluded private school, IEP, and LEP students in that state, and the extent to which these students would have performed differently from students who did participate in the assessment. Because the TSA results in one state are likely to be compared with those in other states, it is also important to consider whether local administrators in the various states applied the criteria for excluding students in a consistent manner, as well as whether there are true differences in the proportions of private school, IEP, and LEP students in the state populations that would affect the comparability of the results.

These various aspects of sample design and implementation are discussed below.

Sampling of Schools

The Panel commissioned a sampling expert to examine the design and execution of the 1990 TSA. His review confirmed that the TSA sample was well designed and conformed to the standards expected for large-scale surveys of this type.[1] The design called for sampling about 100 schools per state. In states that had fewer than 100 schools, all were selected; otherwise, schools were selected with probabilities proportional to their estimated eighth grade enrollments—the larger the school's enrollment, the greater the likelihood that it was included. This ensured that, within each state, each student had an approximately equal chance of being selected into the sample. Schools were sampled from state lists prepared using the National Center for Education Statistics (NCES) Common Core of Data for 1988, but adjustments were made to ensure that schools opened after 1988 also had the opportunity to be sampled.

Sample efficiency, which minimizes the size of the sample required to make accurate estimates for the whole population, was enhanced by stratification—a systematic means for selecting schools within a state that ensures that the resulting sample will reflect the overall composition of the state on certain key variables. For the TSA, stratification was based on urbanicity (central city, suburban, other), percentages of black or Hispanic students, median income, and eighth grade enrollment.

Sampling of Students

Within each selected school, a systematic random sample of eighth grade students was drawn from a list provided by the school. For logistical reasons, these lists were provided to the contractor several months before the actual administration. The resultant time lag meant that some of the students selected had moved to other schools by the date of the assessment; to represent students who had moved, each participating school was instructed to add to its sample a proportion of the eighth graders (if any) who had transferred *into* the school since the original list was drawn.

Analyses of the TSA data suggest that transfer students were, nevertheless, substantially underrepresented in the assessment. According to the contractor's report,

[1] B.D. Spencer, "A Critique of Sampling in the 1990 Trial State Assessment," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

4.9 percent of the students "withdrew" from the sampled schools (i.e., were no longer in the eighth grade at the participating school), but the supplemental sampling procedure added only 2.9 percent to the sample.[2] This difference suggests either that some transfer students were missed through delays in transfer or that some students dropped out of school between the time of their inclusion in the sampling frame, in the fall, and the time of the administration, in February. ***The Panel suggests that in future assessments NCES request more complete information from schools on withdrawals (e.g., date of withdrawal, withdrawal status)*** in order to better understand the dynamics of the transfer process and create a more adequate sampling design for this segment of the population. Procedures for preparing and sampling from the supplemental lists should also be reviewed to ensure that there is no artifactual reason for the apparent underrepresentation of transfer students.

Participation

Recruitment of schools from the sample drawn by the contractor was carried out by state coordinators, individuals designated by their state departments of education to manage the assessment in their states. NCES and its contractors, the Educational Testing Service (ETS) and Westat, worked in a variety of ways to facilitate the efforts of these state coordinators and to maintain their commitment to the TSA process. In addition, participation rates were kept high by providing the state coordinators with matched replacements for schools that refused to participate and by instructing local test administrators to hold makeup sessions in any schools where four or more students were absent from the original assessment.

Overall, 94 percent of the originally sampled schools agreed to participate, but the participation rate was less than 80% in two states: Illinois (78 percent) and Oklahoma (76 percent).[3] Recruitment of replacement schools raised the final overall rate to 97 percent of the planned total, although New York, the state with the lowest final rate, ended with only 86 percent. The average participation rates for the selected students were also generally good, ranging from 92 percent to 96 percent across states, except in the District of Columbia and Oklahoma, where they were 88 percent and 80 percent, respectively.[4]

Statistical analyses did not reveal any major effects of school nonparticipation on the estimates of state mathematics achievement. The effects of student nonparticipation were more difficult to evaluate because machine-readable information was not available on the characteristics of students who were absent or attended makeup

[2] Westat, Inc., *Report on Field Operations and Data Collection Activities, 1990 National and Trial State Assessments* (Rockville, MD: 1990).

[3] School participation rates for these two states may have been low due to the influence of strong local control over schools, combined with a feeling, at the local level, that there is too much testing going on already. However, participation may also have been a function of the recruitment strategies attempted with the original sample of schools, because the rates in both these states were as good or better than most other states after substitution. Final school participation rates were 96 percent for Illinois and 99 percent for Oklahoma.

[4] Student participation rates were low in Oklahoma because state law required that each student being tested present a signed parental consent form. In the District of Columbia, there were only 36 schools with eighth grade students, so testing was done in very large sessions. This circumstance, combined with the high rates of absenteeism in the District, made it unusually difficult for the local administrators to ensure attendance at the assessment sessions.

sessions. (As noted below, NCES expects to collect more detailed information on these students in future assessments.)

Weighting and Adjustments for Nonresponse

Statistical weights are used both to adjust for different selection rates and to address problems in participation. In the computation of state averages, some participants' scores are given more weight than others; in particular, the weight given to a participant's score is determined by how many students he or she represents in the total population. The Panel found that the adjustment methods used in the TSA for unequal selection probabilities were reasonably chosen and appropriately applied.

To adjust for nonresponse, weights for nonparticipating students are reassigned to similar students who did take the test. In the 1990 TSA, matching of absent students with participating students was based on school characteristics, whether or not a student attended a monitored session, and age (less than 13 versus 13 or older). However, other important information that would have enhanced the matching process (particularly individual student information on race, ethnicity, and sex) was not used. The precision of the statistical adjustment for nonresponse was thus somewhat limited. The overall nonresponse rate was low enough in 1990 so as not to create a significant problem in using the results of the assessment. *Nonetheless, the Panel strongly supports the NCES plan to collect and use demographic information about nonresponding students in 1992.*

All estimates based on samples contain a certain amount of inexactitude, referred to as sampling error. The size of the sampling error (and hence the precision of the estimates that can be derived from the assessment results) is largely a function of the size of the sample. The review commissioned by the Panel indicated that the magnitude of the sampling error in the 1990 TSA was reasonable and commensurate with the standards generally expected for a large-scale survey of this type. However, in using TSA data to make estimates for groups or characteristics that are only sparsely represented in the population, the impact of the sampling error became more problematic. This can be seen, for instance, in the statistics presented by race-ethnicity. Of 40 participating states and territories, 38 had samples that were adequate to produce reliable estimates of white students' performance, but only 11 produced performance estimates for Asian Americans and Pacific Islanders that ETS considered sufficiently reliable to publish.[5]

Excluded Students

P.L. 100-297 specifies that all public school students who are selected for the TSA and are capable of participating should be assessed, but it allows for the exclusion of private school students, and IEP or LEP students who are judged incapable of participating meaningfully in the assessment. Nationwide, one in six students belongs to at least one of these three groups (private school, LEP, IEP), and differences among

[5] All of the participating states and territories produced performance estimates for Hispanic students, 31 produced estimates for black students, and 9 produced estimates for American Indians. U.S. Department of Education, National Center for Education Statistics, I.V.S. Mullis et al., *The State of Mathematics Achievement* (Washington, D.C.: 1991), 266.

the states in the proportions of students in these groups could have important effects on state-by-state and state-to-national comparisons. Therefore, the sampling expert commissioned by the Panel examined the effects of exclusions in some detail.[6]

Private School Students. Private school students typically made up the largest excluded group in each state participating in the TSA. Nationally, about 12 percent of eighth grade students are enrolled in private schools, but percentages vary widely across states. In seven states, fewer than five percent of eighth grade students are in private schools; in seven others, more than 18 percent are in private schools. Wyoming and Utah each enroll only about two percent in private schools, whereas Hawaii and the District of Columbia each enroll about 20 percent. Given this wide variability, the Panel is concerned about the effects of private school exclusion on the comparability of state-to-state results.

The effects of this wide variability of private school enrollment across states would not be so problematic, were it not for the fact that private school students tend to perform significantly better than public school students, at least in the lower grades. National NAEP data for 1990 show that the average NAEP mathematics proficiency score was 278 for eighth grade Catholic school students and 274 for eighth grade students from other private schools. Eighth grade public school students, however, scored significantly lower, with an average proficiency of 264.[7] Extrapolating these probable performance differences to the TSA, analyses performed on behalf of the Panel indicated that ranks for some of the participating states would have changed noticeably if private school students had been included.

State NAEP data would better reflect educational achievement and make state results more readily comparable if results for all students, private as well as public, were included. Ultimately, because of the significant variation from state to state in private school enrollment, the exclusion of private school students from the TSA will diminish its utility. The Panel therefore recommends that future authorizations for state NAEP include adequate resources to sample private school students, thereby increasing the comparability of results from one state to another.

Individualized Education Plan (IEP) Students. In ideal testing circumstances, the NAEP assessment would sample equally from all subgroups in the population, including IEP students, to give the best estimates of trends in achievement for all American students. For logistical, ethical, and cost reasons, however, this is not possible for NAEP. The costs of producing different versions, tailored to the needs of various subgroups, would be prohibitive when one takes into account the necessities dictated by such constraints as learning disabilities, communication disorders, mental retardation, emotional problems, and orthopedic impairments.

The sampling design for the TSA allowed for the exclusion of IEP students who were judged incapable of participating meaningfully in the TSA. Local administrators were advised that decisions about the participation of IEP students who were mainstreamed

[6] B.D. Spencer, "Eligibility/Exclusion Issues in the 1990 Trial State Assessment," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

[7] U.S. DoE, NCES, I.V.S. Mullis et al., *The State of Mathematics Achievement*, 90.

at least 50 percent of the time would have to be made by the school's IEP team or equivalent group; they were told also that they should err on the side of inclusion rather than exclusion.[8] IEP students who were tested were identified as such on the administration schedules.

Overall, about four percent of the sampled students were excluded from the 1990 TSA because they had individualized education plans and were evaluated as not able to participate. The percentages varied across states, however, ranging from about two percent in Montana to eight percent in Arkansas. Unfortunately, there was no suitable external data source against which to compare each state's exclusion rate to determine whether this level of state-to-state variation was reasonable. However, the Panel notes that, overall, the TSA *included* roughly twice as many *test takers* who were classified as IEP as did the national NAEP (4.0 percent versus 1.7 percent). This suggests that the local administrators for the TSA may have applied the IEP exclusion rules somewhat differently than the contractor personnel who administered the national NAEP.

Significant variations in the state-by-state exclusions of IEP students were observed in the 1990 TSA that cannot be easily explained. Importantly, the Panel's research shows that differential exclusion rates affect the rankings of the states. Therefore, the Panel recommends that NCES conduct a study designed to evaluate the rationales used by educators for the exclusion of IEP students on the basis of their ability to participate meaningfully in the assessment. This study would result in a better understanding of the differential use of exclusion criteria across states, thereby providing information that would allow states to compare themselves more accurately on NAEP assessments.

For example, existing normative data on both excluded and included IEP students at selected schools could be used to independently evaluate levels of functioning for these students, and persons responsible for the IEP exclusion decisions could be surveyed to obtain more information about their decision processes. Classroom or resource center teachers could also be contacted for additional information on the functional status of included and excluded IEP students.

Limited English Proficiency (LEP) Students. LEP students were excluded from the TSA if they were native speakers of a language other than English, had been enrolled for less than two years in an English-speaking school (not including a bilingual education program), *and* were judged incapable of taking part in the assessment in a meaningful way. As with IEP students, local administrators were told to include doubtful cases in the assessment.

Overall, about one percent of the students in the 1990 TSA sample were excluded because of limited English proficiency. State by state, LEP exclusions ranged from near zero in West Virginia and about two percent in New Jersey, New York, Rhode Island, and Texas, to five percent in California. The differences in exclusion rates due to LEP across states are not surprising, given the ethnic compositions of the states. However, it is possible nonetheless that the exclusion criteria were applied inconsistently across or within states. Paralleling the case of the IEP students, a higher proportion of test takers was found to be classified as LEP in the TSA than in the national NAEP (1.4 percent compared with 0.8 percent).

[8] Westat, Inc., *Summary for All States of Training and Data Collection Activities, Preliminary Results, 1990 Trial State Assessment* (Rockville, MD: 1990).

As noted, in some states, excluding students who do not read English well enough to allow them to be assessed by NAEP means excluding as many as five percent of eighth grade students. Moreover, the rates will be higher when fourth grade students are assessed because the numbers of limited English-speaking students are higher in the earlier grades. *The Panel therefore believes that serious consideration should be given to the development of non-English versions of NAEP for implementation on a trial basis.* To encourage attention to this issue, the Panel is commissioning a study on the feasibility, cost, and advisability of non-English versions of the mathematics assessment as part of its continuing evaluation of the TSA. The Panel will report its findings, with the evaluation of the 1992 trials, to the acting commissioner of the NCES with the purpose of providing them to Congress.

The Implementation of the 1990 TSA

If measures of student performance are to be accurate and fair, all students must receive the same directions, be given the same time limits, and be restricted by the same controls. National NAEP has achieved a high level of standardization by having trained contractor staff administer the assessment. For practical and economic reasons, however, the TSA was administered by local education personnel who had been trained by the contractor. Given this, there was a particular need to examine questions of uniformity in the administration of the first TSA.

Training and Monitoring of Local Administrators

Before the 1990 TSA was administered, teachers, counselors, and other school personnel who were to serve as local administrators were trained by contractor staff in a day-long session that covered pre- and post-assessment activities, as well as conduct of the assessment. The training materials included a videotape explaining the importance of NAEP and its assessment principles, as well as specific exercises to be completed by the local administrators and reviewed by the trainers. In addition, the NAEP contractor employed quality control monitors. A few days before the assessment, these monitors called all local administrators to confirm that they had received their materials and completed preliminary steps such as the distribution of Teacher and Excluded Student questionnaires. The quality control monitors then attended a randomly selected half of the test sessions in each state, where they recorded any observed irregularities *and* intervened to correct these irregularities when they threatened the validity of the assessment. All test sites were candidates for monitoring, and no local administrator knew in advance if a monitor would be present at his or her session.

Session Administration

Irregularities in session administration create especially unfavorable, or especially favorable, testing conditions for particular students. For example, inaccuracies in the timing of the subtests can result in some students having more or less time than

intended to complete certain testing tasks, and inadequate supervision by test administrators or crowded seating can create opportunities for some students to copy from the work of others. Similarly, administrators' errors in reading instructions can create confusion and misunderstanding so that some students do not properly execute the testing tasks, while unsanctioned elaborations on instructions, or ad-lib pep talks, may lead to higher scores.

The Panel's independent analysis of data from monitored and unmonitored sessions indicates that the 1990 TSA administration went very well and that session irregularities that could significantly affect student performance were infrequent. The best evidence for this is that student performance (both mean levels of performance and variability of performance within sessions, between schools, and between states) was essentially the same in monitored and unmonitored sessions.[9]

Among those factors under the control of the local administrator, the irregularities most frequently noted by the quality control monitors were errors in reading the test instruction scripts (additions, omissions, or revisions), and errors in distributing and collecting the calculators between item blocks.[10] Errors in three or more script sections (out of five) were observed in 13 percent of the monitored assessment sessions, and errors in calculator distribution were observed in 11 percent. The monitors also noted the presence of student disruption in seven percent of the monitored sessions.

Efforts to examine the relationships between irregularities noted by the quality control monitors and student performance were somewhat limited by the fact that the monitors intervened to correct problems when they threatened the quality of the assessment, thus attenuating the impact on student scores. Nonetheless, the Panel uncovered some modest relationships between variations in administrative procedures and student performance. Significantly lower student scores were observed in sessions where local administrators made frequent errors reading the instructions and in sessions where the students were disruptive.[11] Fortunately, these and other session administration irregularities tended to be distributed randomly throughout the participating states, with no state as a significant outlier. Thus these irregularities do not diminish the validity of the state-by-state comparisons.

In evaluating the continuing utility of session monitoring in light of these generally very positive findings about the quality of the administration, one must consider that the unmonitored sessions may have gone as well as they did because local administrators prepared more thoroughly in the expectation that their sessions might be monitored. Moreover, the monitored sessions offered the only window into what

[9] D. H. McLaughlin, F.B. Stancavage, J.G. Chambers, E. Hartka, and K. Ercikan, "Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

[10] "Item blocks" were sets of items that always appeared together in the assessment. Under the matrix sampling design used by NAEP, each student attempted three separately timed item blocks. In order to test math skills using both calculators and hand calculations, calculators had to be distributed and picked up between item blocks. Moreover, in order to maximize the use of an expensive resource, the blocks were staggered so that different students would be using the calculators at different times.

[11] With regard to the latter relationship, it should be noted that the direction of the effect is unclear. Disruptions may lower performance or may simply be more likely to occur in schools with large numbers of low achieving students.

went wrong, how frequently it went wrong, and how greatly it affected student performance.

Therefore the Panel concludes that monitoring the administration of the 1990 mathematics assessment provided valuable data and helped to assure uniform testing conditions. As the TSA expands into more states, and different administrators are trained to carry out the testing, the Panel recommends that random monitoring of schools be continued. Uniform procedures for administration are essential for helping assure comparable results across states.

Student Motivation

Because of questions about possible differences in testing conditions or in student motivation to participate, the Panel compared the composite results of the TSA to the composite results obtained for a specially constructed matched subset of students in national NAEP. (The use of the matched subset, which was prepared by the NAEP contractor, was required to compensate for the fact that the sampling frame for the TSA was not the same as that of national NAEP.) Based on these comparisons, students in the TSA were found to have obtained slightly, but reliably, higher scores. Although a definitive explanation for the difference is not possible, the Panel hypothesizes that students participating in state NAEP may have had more motivation to do well than those in national NAEP. The Panel will monitor the 1992 trials closely to see if this potentially important finding is replicated.

In the TSA, quality control monitors recorded instances of local school personnel giving students incentives to participate. The students in nearly one quarter of the monitored sessions had met with school personnel prior to the assessment: in 15 percent of the sessions they had received a pep talk; in nine percent they had been given an incentive for participation such as pizza or ice cream after the assessment, or tickets to school events; and in five percent they had received special instructions on using calculators. Analyses of performance data showed that students who had been given incentives to participate tended to perform *less well* than others. However, this may have been due to the fact that the incentives were not contingent on level of performance, only on participation. It is also possible that these incentives were more likely to be given in schools where student achievement was known to be lower than average.

A separate analysis showed that scores in sessions conducted by principals, vice principals, and school superintendents tended to be *higher* than scores in sessions conducted by teachers or counselors. This may indicate that the perceived status of the test administrator had an effect on results, although an alternative explanation is that administrators may have chosen to participate more often in higher achieving schools.

NAEP is currently a low-stakes assessment. Students do not receive an individual score for their performance, and neither are average scores reported at the local district or school level. These circumstances may somewhat reduce external pressure and intrinsic motivation to perform well, thereby depressing students' performances and yielding underestimates of achievement. The Panel commends efforts by the National Assessment Governing Board (NAGB) and the NCES Technical Review Panel (TRP) to conduct studies that will help to reveal the role that motivation plays in the TSA results.

Feasibility of State NAEP

Finally, in considering the feasibility of the TSA, the Panel has been alert to a concern expressed in P.L. 100-297 that state NAEP might have a negative impact on participation in national NAEP, especially in small states. The Panel also sought to determine the continuing receptivity of the state testing directors to state NAEP as well as their perceptions of the benefits and costs of this testing initiative.

State NAEP's Impact on National NAEP

Although it is true that state NAEP increased the burden on small states to provide sufficient numbers of schools to meet the requirements of both the national and state samples, few of the 37 states that participated in the first trial found this to be a problem. Moreover, the 1990 national NAEP school cooperation rates were similar to the cooperation rates obtained in other recent NAEP administrations, and the 1990 TSA school cooperation rate was higher. (See Table 1.) It is of significance as well that the administration of the 1990 trial did not interfere in any way with the administration of national NAEP. There is, therefore, no indication at this time that a program of state NAEP testing will have a negative impact on the national indicator system. However, the Panel recognizes that this will have to be reevaluated after the 1992 trial in which the testing burden will have increased due to the addition of assessments at grade four in reading and mathematics.

Table 1

School cooperation rates (percent of originally selected schools participating) in recent national and state NAEP assessments

	Grades 3/4	Grades 7/8	Grades 11/12
1983-84 national NAEP	89%	90%	84%
1985-86 national NAEP winter sample	87%	84%	N/A
1985-86 national NAEP spring sample	89%	89%	81%
1990 national NAEP	88%	87%	81%
1990 TSA	N/A	94%	N/A

Sources: U.S. Department of Education, National Center for Education Statistics, A.E. Beaton et al., *Implementing the New Design: The NAEP 1983-84 Technical Report* (Washington, D.C.: 1987), 87. U.S. DoE, NCES, A.E. Beaton et al., *Expanding the New Design: The NAEP 1985-86 Technical Report* (Washington, D.C.: 1988), 71. U.S. DoE, NCES, I.V.S. Mullis, et al., *The State of Mathematics Achievement*, 453. U.S. DoE, NCES, S.L. Koffler et al., *The Technical Report of NAEP's 1990 Trial State Assessment Program* (Washington, D.C.: 1991), 55.

Perspective of the State Testing Directors

State testing directors are knowledgeable observers and important stakeholders in the assessment process. To gain their perspective on the TSA, these directors were surveyed in late 1990 as part of the Panel's study on implementation; at that time, most reported that the assessment went well in their states and nearly three quarters felt that the 1990 trial was likely to be worth the cost to their states. That this generally positive attitude continues is borne out by the fact that there has been an increase in the number of states that will participate in the TSA in 1992.

Some of the state testing directors, however, also noted that they had not been consulted at appropriate points about important policy decisions in the establishment and implementation of the TSA. In its April 1991 interim report to Congress, the Panel commended the efforts of NAGB and NCES in their stated intention to use the Education Information Advisory Committee of the Council of Chief State School Officers (CCSSO) as an arena for providing state testing directors with greater opportunity for policy input. In addition, the Panel proposed that NAEP strengthen its mechanisms for securing state testing directors' input into the state NAEP policy and assessment development process. The Panel continues to endorse the view that NAGB and NCES should provide the organizational structure and resources that insure meaningful input from the state testing directors into state NAEP governance and administration.

3 The Content Validity of the Assessment

Overview

Validity is the most important consideration in test evaluation, because it refers to the appropriateness, meaningfulness, and usefulness of the specific inferences that will be made from the test scores. Because the scores on the NAEP mathematics assessment are the basis for inferences about U.S. students' knowledge of mathematics, the first job of the test developers is to specify adequately the content of mathematics it is important for our students to know. Content validity is then judged by evaluating the degree to which the sample of items in the assessment adequately represents this defined domain of mathematical knowledge and skills. The specification and sampling of content are far from simple tasks, however, because they inevitably involve compromises between the visions of what mathematics literacy should be, and the realities of opportunities to learn in today's classrooms.

The 1990 mathematics assessment, like other recent NAEP assessments, was based on a set of mathematical process and content area objectives, organized into a matrix, or framework, which guided the development of the test items and the assembly of the test forms. The framework and objectives were arrived at through a consensus process involving representatives of various sectors of the educational community and overseen by the Council of Chief State Schools Officers (CCSSO). These objectives were then translated into items by the NAEP contractor, the Educational Testing Service (ETS).

In this chapter, the Panel evaluates the consensus process that produced the 1990 mathematics framework, as well as the validity of the framework and the items generated from it. In addition, the Panel considers the principles that should guide the evolution of future NAEP content objectives and assessment methodologies, and provides suggestions for overcoming potential problems in the development process.

The main findings from the Panel's study of the content validity of the mathematics assessment are:

- *A national consensus process is an appropriate basis for defining NAEP content objectives and frameworks. The consensus process that generated the 1990 mathematics objectives and framework was judged adequate, although not as broadly representative as would have been desirable. The deficiencies were primarily attributable to the short timeline allotted for the activity, and the fact that this was the first time this process for developing consensus had ever been used by NAEP.*

- *The 1990 NAEP mathematics framework and the assessment design principles on which it was built were essentially sound. Further, the framework represented a reasonable compromise between current instructional practices and the standards being put forward by the professional mathematics community at that time. To remain viable, the objectives and frameworks must be viewed as evolutionary entities which are adapted regularly to reflect changing conceptions of mathematics education and measurement.*

- *An analysis of the 1990 framework and Trial State Assessment (TSA) item pool by eight leading mathematics education professionals confirmed that items for the 1990 assessment generally fit the NAEP framework well. Agreement was 90 percent on item classification by mathematics content areas but somewhat lower for classification by mathematics process categories.*

- *Analysis indicates that over half of the 1990 TSA mathematics items reflect one or more of the major themes of the National Council of Teachers of Mathematics (NCTM) Curriculum and Evaluation Standards for School Mathematics in some non-trivial way. A much smaller percentage of the TSA items were judged to fully incorporate the vision of the Standards. It seems likely that the NAEP mathematics assessment will have to evolve to include more difficult items, more diverse items, and more items in total if it is to capture the full range of mathematical abilities identified by the Standards.*

The 1990 Mathematics Objectives and Framework

The National Assessment Planning Project, conducted under the auspices of the Council of Chief State Schools Officers (CCSSO), formulated a development process for the 1990 mathematics objectives that was consonant with the language of the authorizing NAEP legislation, which stated that "each learning area assessment shall have goal statements devised through a national consensus approach, providing for active participation of teachers, curriculum specialists, subject matter specialists, local school administrators, parents, and members of the general public."[1] In addition, in anticipation of the first reporting of state-level results, the Planning Project determined to expand the development process to ensure careful attention to the formal mathematics objectives of the states and of a sampling of local districts, as well as to the informal opinions of practitioners at the state and local levels.

To guide its efforts, the Planning Project assembled a Steering Committee of policymakers, practitioners, and citizens who had been nominated by 18 national organizations. It also convened a Mathematics Objectives Committee made up of a teacher, a school administrator, and mathematics education specialists from various states, as well as mathematicians, parents, and citizens. The draft objectives, accompanied by a set of sample questions, were distributed at various stages in their development for review by the mathematics supervisors in each of the 50 states, 25

[1] Public Law 98-511, Section 405 (E), which authorized NAEP through June 30, 1988.

other mathematics educators and scholars, the NAEP Assessment Policy Committee and item development panel, and other interested groups. In addition, the state mathematics supervisors were encouraged to convene review groups of their own to provide a wider perspective on the adequacy of the proposed objectives.[2]

The framework that emerged from this process can be represented as a matrix of five mathematics content areas crossed with three dimensions of mathematical process, shown in Figure 1. The figure also indicates the proportions of items in the eighth grade assessment to be drawn from each content or process area. (The proportions were set separately for each grade so as to accommodate age appropriate assessment of knowledge and skills.)

Figure 1

Framework for the 1990 mathematics assessment showing the percentage distribution of items for the eighth grade by content area and mathematical process

Content area	Mathematical process			
	Conceptual understanding	Procedural knowledge	Problem Solving	Percent of items
Numbers & operations				30%
Measurement				15%
Geometry				20%
Data analysis, statistics, & probability				15%
Algebra & functions				20%
Percent of items	40%	30%	30%	100%

Educational Testing Service, *Mathematics Objectives*, 14.

The framework was based on a number of assessment design principles that the Mathematics Objectives Committee articulated:

- Although no assessment can measure everything, the assessment should nevertheless reflect many of the states' curricular emphases and objectives; be inclusive of what various scholars, practitioners, and interested citizens believe should be in the curriculum; and maintain some of the content of prior assessments to allow reporting of trends in performance.

- The assessment should encourage curriculum development by going beyond the "least common denominator" of topics and skills that appear in the objectives of every state. Further, it should not be bound by the

[2] Educational Testing Service, *Mathematics Objectives: 1990 Assessment* (Princeton, NJ: 1988), 6-8.

least common denominator of student preparation but should include challenging questions at each grade level.

- The assessment design should be sensitive to the concern that its objectives not be used to steer instruction toward one particular pedagogical or philosophical viewpoint to the exclusion of others that are widely held.

- Although a matrix is useful to indicate the intended content and distribution of items, the developers acknowledge that the application of mathematics is far more holistic than a matrix implies, and that the mathematics content areas and process categories described in the objectives are not discrete or mutually exclusive, but rather highly integrated.[3]

The Adequacy of the Consensus Process and the Framework

A study commissioned by the Panel began its evaluation of the content validity of the 1990 mathematics assessment with a retrospective review of the adequacy of the consensus process carried out by CCSSO. The study directors examined archival records and other written materials of the Planning Project and conducted interviews with ten individuals who had participated in the state-level consensus activities and 18 individuals who had been involved at the CCSSO level of the planning process.[4]

Based on this study, the Panel concludes that CCSSO made a "good-faith" effort to include the relevant constituencies. The major committees of the Planning Project included members from a broad array of organizations, not only those specifically interested in mathematics education, but those interested in education more generally. Members of the major national organizations associated with mathematics education, such as the National Council of Teachers of Mathematics (NCTM) and the Mathematical Sciences Education Board (MSEB), were included on major committees appointed by the Planning Project, although not as direct representatives of these organizations. Publications of these organizations were also consulted, and it is clear that their perspectives were influential in the consensus process.

Based on the review of records and interviews with 10 individuals who participated at the state level (eight selected at random and two self-nominated), it would appear that the state consensus process, although adequate, was neither as comprehensive in its mandate nor as broad in its representation as would have been desirable. On the positive side, at least 39 states actively participated in the process by providing feedback on the draft objectives and framework to the Planning Project. The negatives included the lack of systematic representation in the groups assembled at the state level, the lack of opportunity for *inter*-state consensus-building (each state committee worked in isolation), and the somewhat limited nature of the judgments required of the state committees, all of which may have reduced the possibility of

[3] Ibid., 9-11.

[4] E.A. Silver, P.A. Kenney, and L. Salmon-Cox, "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

meaningful debate or disagreement. For example, the study found that the states made very different decisions about how to assemble groups for the review process; some used groups formed for other purposes, some formed ad hoc groups, and some handled the entire review by mail and phone. Moreover, the participation of some critical groups—including the teachers in at least one state—was precluded due to the short timeline. With regard to the scope of the task, the state committees were given approximately two weeks to review (a) the reasonableness of the content-by-process organization of the draft framework, (b) the inclusiveness of the content areas to be included in the assessment, (c) the appropriateness of the grade level designations, (d) the proposed distribution of the percentages of items in the process and content categories, and (e) the appropriateness of the item formats (e.g., multiple-choice, open-ended). Although encouraged to do so, the states were not required to generate their own frameworks or to give serious consideration to the question of whether a two-dimensional framework with discrete cells for each item type was the best model for structuring test content.

Most of the noted deficiencies were attributable to the short timeline and the novelty of the consensus process. Almost without exception, those interviewed said that the time allotted to the various phases of the consensus process was too short for proper implementation.

Despite these limitations, most of the respondents in the Panel's study felt that the final framework for the 1990 NAEP assessment represented a reasonable compromise between current instructional practices and the new views of mathematical competence being articulated by professional organizations. (Two state mathematics specialists who were interviewed did not agree with this judgment, however. One felt that the framework did not push far enough beyond current practice and the other felt that it had gone too far.) Agreement on the value of a consensus process in determining NAEP frameworks was unanimous.

In addition to the interviews with individuals who participated in the objectives consensus process, the Panel commissioned an independent analysis of the framework and item pool by eight mathematics education experts representing different constituencies and perspectives. Based on the combined evidence from these sources, *the Panel concludes that the 1990 NAEP mathematics framework and the assessment design principles on which it was built were essentially sound.* The designers of the 1990 framework appear to have captured the spirit of change in mathematics education during the late 1980s while also preserving traditional elements; they thus created a reasonable compromise between the visions of educational reformers and the realities of educational practice for that time period.

There was also, however, virtual unanimity in the view that *assessment frameworks must be viewed as evolutionary entities, rather than structures that are fixed for long periods of time.* The framework developed for 1990 must therefore be allowed to evolve to reflect changing conceptions of mathematics education in the nation.

The Fit of the Items to the NAEP Framework

After determining the appropriateness of the mathematics content domain specified by the framework, as well as the consensus process upon which it was based, the next

step in evaluating the content validity of the 1990 NAEP mathematics assessment was to investigate whether the specified domain was adequately represented by the assessment items. To this purpose, copies of all items used in the Trial State Assessment (TSA) were obtained from the contractor and reviewed, during a two-day meeting, by the eight mathematics education professionals mentioned above.

After extensive discussion of the categories and meaning of terms, six of the eight experts classified each of the items used in the TSA according to the content areas and process categories specified by the framework.[5] The raters worked in pairs, with each pair classifying items from three of the seven TSA item blocks. ("Item blocks" were sets of items that always appeared together in the assessment. Under the matrix sampling design used by NAEP, each student attempted three item blocks.) The 23 items in the trend block, a block of items administered in both 1986 and 1990 and then used to link performance across assessments, were classified by all pairs to provide a measure of interrater agreement. The level of interrater agreement achieved was 78 percent.

With regard to the five mathematics content areas, there was a high level of agreement between the classifications of the expert raters and the classifications provided by the NAEP contractor. Disagreements occurred on only 14 of the 137 items used in the 1990 TSA, yielding an overall agreement rate of 90 percent. For the separate content areas, agreement ranged from 85 percent for "geometry" and "algebra and functions" to 95 percent for "measurement" and "data analysis, statistics, and probability" (See Table 2). With regard to the mathematics process categories, agreement was also high for the categories of "procedural knowledge" (88 percent) and "problem solving" (76 percent). Only 51 percent of the items classified as "conceptual understanding" by NAEP, however, were also classified in this manner by the expert raters (See Table 3). In most cases, the disputed items were classified as "procedural knowledge" by the expert raters.

Table 2

Content classification of 1990 TSA items: Cross-classified by NAEP and by expert raters

	Expert raters					
NAEP	Numbers & operations	Measurement	Geometry	Data analysis, statistics & probability	Algebra & functions	Total items as judged by NAEP
Numbers & operations	40 (91%)	—	—	1	3	44
Measurement	—	20 (95%)	—	—	1	21
Geometry	1	3	22 (85%)	—	—	26
Data analysis...	1	—	—	18 (95%)	—	19
Algebra & functions	2	2	—	—	23 (85%)	27
Total items as judged by expert raters	45	25	22	19	26	137

Note: Diagonal (percentage agreement) entries based on NAEP (row) totals.

[5] To avoid any appearance of a possible conflict of interest, two experts who had prior involvement in the process of creating the 1990 NAEP framework or in the item development process did not participate in this aspect of the validity study.

Table 3

Process classification of 1990 TSA items: Cross-classified by NAEP and by expert raters

NAEP	Expert raters			Total items as judged by NAEP
	Conceptual understanding	Procedural knowledge	Problem solving	
Conceptual understanding	29 (51%)	18	10	57
Procedural knowledge	1	38 (88%)	4	43
Problem solving	3	6	28 (76%)	37
Total items as judged by expert raters	33	62	42	137

Note: Diagonal (percentage agreement) entries based on NAEP (row) totals.

Ambiguities in the definitions of the mathematical process categories were thought to have contributed to these disagreements. For example, "conceptual understanding" is defined by the following paragraph in the *Mathematics Objectives:* "Students demonstrate conceptual understanding in mathematics when they provide evidence that they can recognize, label, and generate examples and counterexamples of concepts; can use and interrelate models, diagrams, and varied representations of concepts; can identify and apply principles; know and can apply facts and definitions; can compare, contrast, and integrate related concepts and principles; can recognize, interpret, and apply the signs, symbols, and terms used to represent concepts; and can interpret the assumptions and relations involving concepts in mathematical settings. Such understandings are essential to performing procedures in a meaningful way and applying them in problem-solving situations."[6] It may also be that the abilities themselves are so interrelated that they cannot be assessed in isolation. For example, an item that asks "What number is a common factor of 10 and 15?" was classified by NAEP as "conceptual understanding" and by the expert raters as "procedural knowledge." In fact, to answer this question correctly one must have both conceptual understanding (of the existence and nature of a common factor) *and* procedural knowledge (of the process used to find such a factor). These findings underscore the caution voiced by the designers of the 1990 mathematics framework that the mathematics content areas and process categories described in the objectives are not discrete or mutually exclusive, but highly integrated in some cases.

To summarize, the fit of items from the 1990 assessment to the NAEP process framework was somewhat problematic, especially with regard to the conceptual

[6] Educational Testing Service, *Mathematics Objectives*, 16.

understanding category. By contrast, *the Panel concludes that the assessment provides a valid reflection of the NAEP content framework.*

Implications for Future Assessments

Because of widespread consensus that the content and form of the NAEP assessments must continue to evolve as American education evolves, the work of the Panel went beyond analysis of the content validity of the 1990 mathematics assessment to consider how the content and the assessment methodologies of the NAEP assessments might evolve and how the process of evolution might be directed.

The Influence of the NCTM Standards

When the specifications for the 1990 NAEP assessment were being prepared, the NCTM *Curriculum and Evaluation Standards for School Mathematics* were still in draft form. Subsequently, the *Standards* have been finalized and widely endorsed as a significant statement on what mathematics should be taught in the schools.[7] The Standards emphasize four central themes at all grade levels: mathematics as problem solving, mathematics as communication, mathematics as reasoning, and mathematical connections. These themes are further explicated in Figure 2, as they apply to grades five through eight.

Figure 2

Four central themes in the NCTM Standards, as applied at grades five through eight

Mathematics as problem solving

Problem solving is the process by which students experience the power and usefulness of mathematics in the world around them. In the classroom, problem situations can establish a "need to know" and foster the motivation for the development of concepts.

The nonroutine problem situations envisioned in these standards are much broader in scope and substance than isolated puzzle problems. They are also very different from traditional word problems, which provide contexts for using particular formulas or algorithms but do not offer opportunities for true problem solving. Situations that allow students to experience problems with "messy" numbers or too much or not enough information or that have multiple solutions, each with different consequences, will better prepare them to solve problems they are likely to encounter in their daily lives.

[7] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: 1989).

Figure 2, continued

Mathematics as communication

The use of mathematics in other disciplines has increased dramatically, largely because of its power to represent and communicate ideas concisely. The ability to read, write, listen, think creatively, and communicate about problems will develop and deepen students' understanding of mathematics.

Teachers foster communication in mathematics by asking questions or posing problem situations that actively engage students, including situations that encourage students to create problems themselves. Small-group work, large-group discussions, and the presentation of individual and group reports create an environment in which students can practice and refine their growing ability to communicate mathematical thought processes and strategies.

Mathematics as reasoning

Reasoning is fundamental to the knowing and doing of mathematics. Although most disciplines have standards of evaluation by which new theories or discoveries are judged, nowhere are these standards as explicit and well formulated as they are in mathematics. Conjecturing and demonstrating the logical validity of conjectures are the essence of the creative act of doing mathematics.

The seeds of logical thinking are planted as students learn to describe objects or processes accurately and to elaborate their properties, similarities, differences, and relationships. Students should be encouraged to explain their reasoning in their own words. Listening to their peers and their teacher describe other strategies helps students refine their thoughts and the language they use to express their thoughts.

Mathematical connections

The intent of this standard is to help students broaden their perspective, to view mathematics as an integrated whole rather than as an isolated set of topics, and to acknowledge its relevance and usefulness both in and out of school. Students should have many opportunities to observe the interaction of mathematics with other school subjects and with everyday society. This integration of mathematics into contexts that give its symbols and processes practical meaning is an overarching goal of all the standards.

Persistent attention to recognizing and drawing connections among topics will instill in students an expectation that the ideas they learn are useful in solving other problems and exploring other mathematical concepts. For example, a knowledge of area can help them in understanding the operations on fractions, representing statistical data, solving proportion problems, finding factors and probabilities, and exploring the meanings of algebraic expressions. Different representations of problems serve as different lenses through which students interpret the problems and the solutions. If students are to become mathematically powerful, they must be flexible enough to approach situations in a variety of ways and recognize the relationships among different points of view.

National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics*, 75-86.

As state and local mathematics programs come to embody more and more of the principles set forth in the *Standards*, it seems clear that future versions of the NAEP mathematics assessments will also evolve in this direction. To gauge the distance that will have to be covered to bring NAEP fully in line with these NCTM *Standards*, the expert reviewers assembled on behalf of the Panel examined the extent of the fit between the 1990 TSA item pool and the *Standards*.

Aside from the trend block (which maintains earlier item types by design), the expert raters found that nearly 60 percent of the TSA items related in some non-trivial way to these four major themes of the NCTM *Standards*. A much smaller percentage of the TSA items were judged to fully incorporate the vision of the *Standards*. The strongest and most frequent relationships were to the themes of reasoning and problem solving, while relationships to the theme of connections were virtually nonexistent. (The review did not, however, encompass the three tape-paced item blocks that were used only in the national assessment. It seems likely that these blocks would have contained a larger proportion of items aligned with the *Standards* because they were largely focussed on the measurement of "higher order thinking" and estimation skills.)

In considering the extent to which the 1990 assessment reflected the spirit of the NCTM *Standards*, the raters also expressed concern regarding the high proportion of multiple-choice items in the TSA item blocks. When one includes the trend block, there were about three times as many multiple-choice as open-response items, and only a fraction of the open-response items went beyond "fill in the blank" to truly expand the range or level of skills being tested. Another area of concern was the set of items designed to require the use of a calculator. Because appropriate use of calculators is a curricular goal supported by the *Standards*, these items might have been expected to be among the more forward-looking parts of the assessment. Instead, they were judged to involve only standard, uninteresting computations.

Evidence from the achievement-levels-setting process for the 1990 NAEP combines with the experience of these expert raters to suggest that the NAEP mathematics assessment will have to evolve to include more difficult items, more diverse items, and more items in total if it is to capture the full range of mathematical abilities envisioned by the NCTM *Standards*. The Panel commends the decision of the National Assessment Governing Board (NAGB) to address all three of these issues in the 1992 mathematics assessment.

The Need for Balance in NAEP Content and Assessment Methodologies

Although the Panel anticipates that the NAEP mathematics content objectives and item formats will continue to evolve in the direction of the NCTM *Standards*, the need to balance this vision against the realities of current classroom practices will continue for some time to come. The problem of selecting among competing priorities will be even more intense in other subject areas where the level of national consensus reflected in the NCTM *Standards* does not yet exist. In keeping with NAEP's paramount role as a monitor of educational progress, the Panel suggests that NAGB be guided by the following principles when developing future cycles of NAEP assessments.

NAEP must embody the best thinking about aspirations for what should be taught in the future and how it should be assessed, at the same time that it reflects the best of current instructional content and assessment methodologies. On the one hand, to rely only on the content and assessment methodologies being advanced by the leaders in relevant fields runs the risk that the assessment will be reduced to tracking the fads and fashions that often dominate fields of study. On the other hand, if NAEP merely reflects present or recent past practices, it fails in its potential to monitor the results of curricular and instructional reform and, more generally, will be incapable of measuring the effects of educational innovations. It is critical that NAEP provide comprehensive coverage of content and assessment methodologies, creating an appropriate balance between established practice and cutting-edge advances in education.

Currently many educators are questioning the limitations of the usual multiple-choice questions for assessing student achievement and are advocating the use of newer technologies that require students to construct extended responses or perform other complex tasks. The 1990 mathematics assessment included some open-response items, and plans for the 1992 and 1994 assessments will greatly increase the number and diversity of extended-response and performance items. The Panel supports these design decisions, because they will result in more comprehensive assessment of the levels and dimensions of student ability. However, although the nation is moving toward more authentic and instructionally integrated assessments, the Panel believes that it is also important to retain a role for multiple-choice and short-answer items on a monitoring assessment such as NAEP, given their known validity, reliability, and cost-effectiveness.

In summary, it is important that NAEP's content be comprehensive, reflecting the most forward-looking, modern approaches, at the same time that it reflects the best of current practice. The Panel applauds the increased testing of higher order thinking skills and challenging concepts in the 1990 TSA and recommends that all NAEP content frameworks evolve in this direction. Similar moves should be made to incorporate cutting-edge assessment techniques, at the same time retaining the best of more traditional methodologies. For future mathematics assessments, the Panel recommends including items that assess *all* the themes identified in the NCTM *Standards* including innovative items that require extended written responses.

Directing the NAEP Development Effort

A common theme emerging from almost every discussion about the 1990 mathematics assessment was the lack of time to complete the complex tasks that were required. Timelines for framework development and review were cited as too short, as were the timelines for item development, review, and field testing. A project on such a tight schedule is bound not only to be conservative in its approach but also to depend too extensively on past practices and existing structures.

The Panel's study of the development of the 1990 mathematics assessment makes it clear that planning must begin well in advance for future assessments so that adequate

time can be allowed to develop and test exemplary assessment tasks and to find ways to measure constructs not well covered by today's assessment methodologies. Especially during times of reform efforts, sufficient time is needed to step back from development processes in order to reflect on both the activities and the products.

More recent NAEP development efforts have continued to suffer from overly tight timelines. For example, CCSSO had less than four months to create the 1992 reading framework and only seven months to create the 1994 science framework. *One-year-at-a-time reauthorizations and extensions just prior to the scheduled field tests are not adequate. Congress should authorize NAEP assessments several years prior to the scheduled administration in order to resolve this ongoing problem.*

To support thoughtful and appropriate innovation, it is also important that knowledgeable persons from the relevant subject areas be included at all steps of the NAEP development process. A strong effort was made to do this during the design of the framework for the 1990 NAEP, and some effort was made to do this during item development and review, but the lack of time for reflection and careful consideration of issues was a limiting factor to the value of such participation. Moreover, although a large number of different groups was involved in the process—on the CCSSO and Educational Testing Service (ETS) committees and as consultants—there was no single monitoring group with deep knowledge of the subject matter of the test to oversee the entire process.

The development of assessments in the various NAEP content areas is a multiphase process, involving numerous technical and policy groups. As a result, the process is subject to fragmentation. The Panel recommends that a mechanism be devised to ensure continuity throughout, beginning with the development of the framework and continuing through the various stages of item specifications, item writing, item scoring, and reporting of the results. This mechanism might be a set of oversight committees, one for each subject matter area, composed of both NAGB members and experts from the subject area communities. These committees would help to ensure effective and on-going communication with those most knowledgeable in the NAEP content areas.

4 Scoring and Interpretation

Overview

The NAEP assessment scores are reported on a scale ranging from 0 to 500. The National Center for Education Statistics (NCES) then interprets the scores using sample anchor items and anchor levels set at 200, 250, 300, and 350 on the scale. NCES also provides qualitative descriptions of performance at each of these levels.

Average NAEP scale scores and the percentages of students scoring at the various anchor levels were the primary reporting devices used up to and including the June 1991 release of the results of the 1990 assessment. In September 1991, the National Assessment Governing Board (NAGB) released the 1990 data in a new format—achievement levels. The achievement levels, which were designed to help states measure their progress toward meeting the Administration's National Education Goals, are statements of how basic, proficient, and advanced students at grades 4, 8, and 12 should perform. This is in contrast to anchor levels which provide descriptions of specific performance levels along the achievement continuum, but are not based on judgments of how much students should know or be able to do. NAGB has indicated its intention to use achievement levels as the primary reporting mechanism for NAEP results in all future assessments.

In light of NAGB's decision, the Panel carefully examined the development of achievement levels for the 1990 mathematics assessment, as well as the validity of their interpretation. The Panel is currently studying whether educators and policymakers find achievement levels more useful than the previously used anchor levels. As a result of its examination so far, the Panel finds:

- *Prior to adopting the achievement levels on a permanent basis, research is necessary to determine whether the new achievement levels convey meaning above and beyond that conveyed by the anchor levels.*

- *Because of technical concerns about the achievement levels, data from the 1990 assessment should not be used as a baseline against which to compare future achievement level results unless there is evidence to justify their use.*

The Use of Scale Scores, Anchor Items, and Anchor Levels

The 1990 trial state NAEP assessed mathematics proficiency in five areas: numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions. Scores reflected performance in each of these five content areas as well as for overall mathematics proficiency. The overall score was based on a weighted combination of the five content subscores.

Like many other indices we encounter in everyday life, the proficiency scale on which the NAEP results are reported has no natural nor inherent meaning: It ranges from 0 to 500. The NAEP scale was established to span all grade and age groups. In base years (1990 was a base year) the scale has a mean of 250 and a standard deviation of 50. The same scale is used for several successive NAEP administrations so meaningful comparisons across years can be made. If the meaning of a NAEP score of 350 is evident to someone, it is only because that person has had considerable experience with the NAEP scales—as is the case for understanding the Fahrenheit scale, the Dow-Jones index, blood pressure readings, and the like. Because NAEP scales are not well known to most Americans, they have little meaning for them.

To give meaning to the NAEP proficiency scale, the National Center for Education Statistics (NCES) anchors it by providing descriptions of student performance at various points along the scale, called anchor levels.[1] Content specialists review sets of items that discriminate between adjacent pairs of anchor levels and prepare written descriptions of the knowledge, skills, and reasoning abilities required to answer the items correctly.[2] The knowledge, skills, and reasoning abilities that characterize anchor items at level 250, for example, are likely to be mastered by students who score at or near that level, but less likely to be within the competence of students who score at the next lower anchor level (level 200). A paragraph description of the skills that characterize each anchor point is included in the NAEP mathematics reports.

Percentages of students scoring at each of the defined anchor levels are reported in addition to mean levels of performance on the 500-point NAEP proficiency scale. Because few students score below 100 or above 400 on the scale, the anchor levels begin at Level 150 and end at Level 350. The 500-point NAEP scale spans grades 4, 8, and 12. Few students in grade eight scored below the 200 level. Therefore, four levels of performance were reported for grade eight mathematics for the 1990 assessment, each separated by 50 points: Level 200—Beginning Skills and Understanding; Level 250—Basic Operations and Beginning Problem Solving; Level 300—Moderately Complex Procedures and Reasoning; and Level 350—Multi-Step Problem Solving and Algebra.

[1] In previous NAEP cycles, these "anchor levels" were called "proficiency levels" or "levels or proficiency."

[2] An item is considered to discriminate between adjacent anchor levels if it is answered correctly by at least 65 percent of the students who score at the higher level and by no more than 50 percent of the students who score at the lower level. In addition, there must be a spread of at least 30 points between the percentage of students who answer the item correctly in the two adjacent anchor levels.

Achievement Levels

NCES has used average NAEP proficiency scale scores and the percentages of students achieving at each of the NAEP anchor levels as its two main ways of reporting the results of NAEP assessments since the 1984 assessment. The results are reported not only for the nation as a whole, but also by categories such as region, race-ethnicity, and sex. In May 1990, NAGB suggested a new method for reporting of results of the 1990 mathematics assessment—achievement levels. NAGB then developed a set of policy guidelines for setting achievement levels for the nation, citing its statutory responsibility for identifying appropriate achievement goals for the nation's school children. Although the May 1990 NAGB document does not describe the achievement levels as standards per se, more recent ones do. The September 1991 report on achievement levels results in *The Levels of Mathematics Achievement* states:

> An alternative to this approach [the anchor levels] involves the use of external standards to interpret the NAEP results. In the Board's view, this approach is far preferable. Descriptions of what students should be able to do, based on a set of established criteria, provide an important perspective on student performance. Student performance can then be evaluated not only in relation to other students, but also in light of standards created by experienced educators, policymakers, and informed judges. [3]

The Educational Standard-Setting Movement in the United States

NAGB set achievement levels for the country during a standard-setting movement that gained substantial momentum over the past two years. Many educators have talked in recent years about the desirability of having national education standards, but few believed it possible to establish standards in the United States, given this nation's decentralized educational system. However, as reported in the October 23, 1991, issue of *Education Week*, there has been "a veritable explosion in the number of groups promoting the concept."[4]

Standards are of particular prominence at the national policy level. The Administration's *America 2000* education strategy calls for the development of a set of national examinations—the American Achievement Tests—to measure the performance of the country's 4th, 8th, and 12th graders. In June 1991, Congress, working with the governors and the Administration, boosted the momentum for the development of national standards by creating the National Council on Education Standards and Testing (NCEST). The Council was charged to examine the "desirability and feasibility" of developing a set of national educational standards and assessments for the nation. Its report to Congress was delivered in January 1992. The Council

[3] National Assessment Governing Board, M.L. Bourque and H.H. Garrison, *The Levels of Mathematics Achievement, Initial Performance Standards for the 1990 NAEP Mathematics Assessment, Volume I: National and State Summaries* (Washington, D.C.: 1991), 2-3.

[4] L. Olson, "'Confusing' Array of Players Charts Course Toward National Standards," *Education Week* 11(8), (October 23 1991): 1.

took a historic step by recommending the establishment of national content and performance standards.[5] The Council also recommended that the National Education Goals Panel (NEGP) appoint a new National Education Standards and Assessment Council (NESAC).[6] NESAC would coordinate the development of the national standards and assessments as well as establishing guidelines and criteria for their development. NESAC would then certify the new standards to make certain they are "world class" standards, subject to review by the reconfigured NEGP. Finally, NCEST recommended that NAEP should function as the overall program or system monitor.

What Are Standards?

In the field of education, standards most frequently refer to either content standards or performance standards, but the distinction is often lost in the course of debate and discussion.[7]

Content Standards. *Content standards* are statements about the content goals for learning and instruction. NCEST, in *Toward Higher Standards,* states: "Content standards...define the knowledge, skills, and other necessary understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter."[8] Perhaps the best example of a set of content standards is the National Council of Teachers of Mathematics' (NCTM) *Curriculum and Evaluation Standards for School Mathematics* (1989) discussed in Chapter 3. These *Standards,* developed by a nationwide group of classroom teachers, supervisors, educational researchers, teacher educators, and college-level mathematicians, articulate the meaning of mathematics literacy in today's highly technological society. NCTM *Standards* place heavy emphasis on the need for a mathematics curriculum that deals with problem solving, the communication of mathematical ideas, critical reasoning, and connections of ideas and procedures both within mathematics and across other content areas. Following on the work done by NCTM in the field of mathematics, NCEST and others have called for the development of content standards in the areas of English, history, geography, and science as well.

Performance Standards. A second classification of standards is *performance standards*. Performance standards differ from content standards in that they describe levels of desired competence or performance after the content standards are developed. Performance standards may be set to measure single or multiple levels of competency. Single levels yield simple pass or no pass results. Multiple levels measure performance against several standards such as pass with honors, pass, or fail.

Performance standards necessarily depend on judgment, and judges often have different views about the performances that indicate an adequate—or passing—level

[5] National Council on Education, Standards, and Testing, *Toward High Standards: A Report to Congress, the Secretary of Education, and the National Goals Panel* (Washington, D.C.: 1991), 2.

[6] NCEST also recommended that NEGP be reconfigured to give voting power to the Congressional members on the Panel.

[7] NCEST also defines and discusses school delivery and system standards. Because neither are relevant for the Panel's evaluation of the NAGB achievement levels, they are not treated further in this document.

[8] *Ibid*., 14

of mastery or understanding. Even if teachers used the same cutting points on a percentage scale to assign grades, the standards of different teachers would not necessarily be comparable because teachers are likely to differ in both the difficulty of the tasks they use for grading and the stringency with which they award points. Furthermore, even seemingly absolute standards are influenced by normative understandings. Although we rarely set standards so high that no one can attain them, we often set the standards beyond the status quo; they can be thought of as aspirations for performance.

How NAGB Set Achievement Levels

NAGB established three achievement levels: basic, proficient, and advanced. A panel of judges consisting of educators, scholars, business leaders, and other qualified specialists convened in Vermont in August 1990. Using a modified Angoff procedure,[9] one of several widely used item-judging procedures, the judges set levels for certification examinations through the following procedure. After reviewing definitions for basic, proficient, and advanced achievement levels, judges formed a mental impression of what a student at a given achievement level in a given grade should be able to do. They then reviewed the entire pool of items used in the 1990 NAEP mathematics assessment and, for each item, independently judged the probability that a basic, proficient, or advanced student would answer it correctly. They repeated this procedure a second time using descriptions of actual student performance on each of the items. After discussing the results of cycles one and two, the judges repeated the procedure a third time. Fourth and fifth cycles of ratings were held a month later in Washington, DC.[10]

NAGB used a procedure that was devised primarily to set standards for professional licensing and certification and in the establishment of minimum competency standards, that is, in situations requiring a single standard. While it is difficult to get agreement among judges about where to set a *single* cutpoint, NAGB asked the judges to use the procedure to establish *three* cutpoints—a far more demanding task. A further complication was that the item pool was chosen *prior* to NAGB's decision to set achievement levels. As a consequence, some judges complained that the item pool did not contain a sufficient number of items at the advanced levels for the 8th and 12th grades.[11]

A number of parties including, but not limited to, NAGB's own evaluation team, the NCES Technical Review Panel, the Council of Chief State School Officers (CCSSO), and this Panel, closely scrutinized the effort to set achievement levels. Based on advice from its evaluation team and other technical advisors and from testimony at two public hearings, NAGB decided that the standards set during the initial trial in late summer and fall of 1990 were not ready for public release. Therefore, additional achievement-level-setting sessions were conducted in four regions of the country in

[9] W.H. Angoff, "Scales, Norms, and Equivalent Scores." In *Educational Measurement*, ed. R. L. Thorndike (Washington, D.C.: American Council of Education, 1971): 508-600.

[10] NAGB, Bourque and Garrison, *The Levels of Mathematics Achievement*, 10.

[11] D.L. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-1991 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (August 23, 1991), 19.

the spring of 1991. The results of the latter effort were eventually used to establish the 1990 NAEP mathematics achievement levels of basic, proficient, and advanced for students at grades 4, 8, and 12.

The adequacy of the achievement levels has been a source of considerable controversy. The NAGB evaluation team concluded that (a) the achievement levels set for the 1990 NAEP assessment should not be used as a "baseline or benchmark" against which to compare future achievement level results and (b) the procedures used to set the levels should not be treated as a model for future NAEP assessments. Contrary to the report of its evaluation team, NAGB adopted the achievement levels. At the request of Congress, this matter is being reviewed by the General Accounting Office. It is the Panel's understanding, however, that the achievement levels data released on September 30, 1991, *will not* be used by NAGB as a baseline against which to compare future achievement level results *unless* additional evidence is garnered to warrant such a use. *The Panel strongly endorses this position.*

The Need to Set Standards in a Relevant Context

The Panel believes that education standards, if they are to be defensible, should *be set in the context of both content standards and classroom instruction*. Developing standards outside this context will very likely result in a set of performance standards with no clear meaning. To see how this can happen, consider NAGB's definition of advanced performance:

> The higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade, the advanced level shows readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams.[12]

Notice that the definition gives little specificity about the academic performances required at grade 12 and none at grades 4 or 8. For example, the basic definition of the advanced level for 12th graders states that it signifies "readiness for rigorous college work." Yet *there are no completed, ongoing, or planned studies to verify that the performance criteria—skills and knowledge—judged necessary for an advanced-level 12th grader are actually related to performance at the college level.* In other words, there is no evidence that the achievement levels predict the criteria described. Furthermore, except for 12th grade mathematics, the criteria which the achievement levels claim to predict were not identified.

From the Panel's perspective, NAGB's decision to move from the use of the anchor levels to the use of achievement levels raises two important questions. First, are the achievement levels an appropriate set of education standards for the nation's 4th, 8th, and 12th graders? Second, are the new achievement levels more useful for reporting than the NAEP scale scores and the anchor levels?

[12] NAGB, Bourque and Garrison, *The Levels of Mathematics Achievement*, 5.

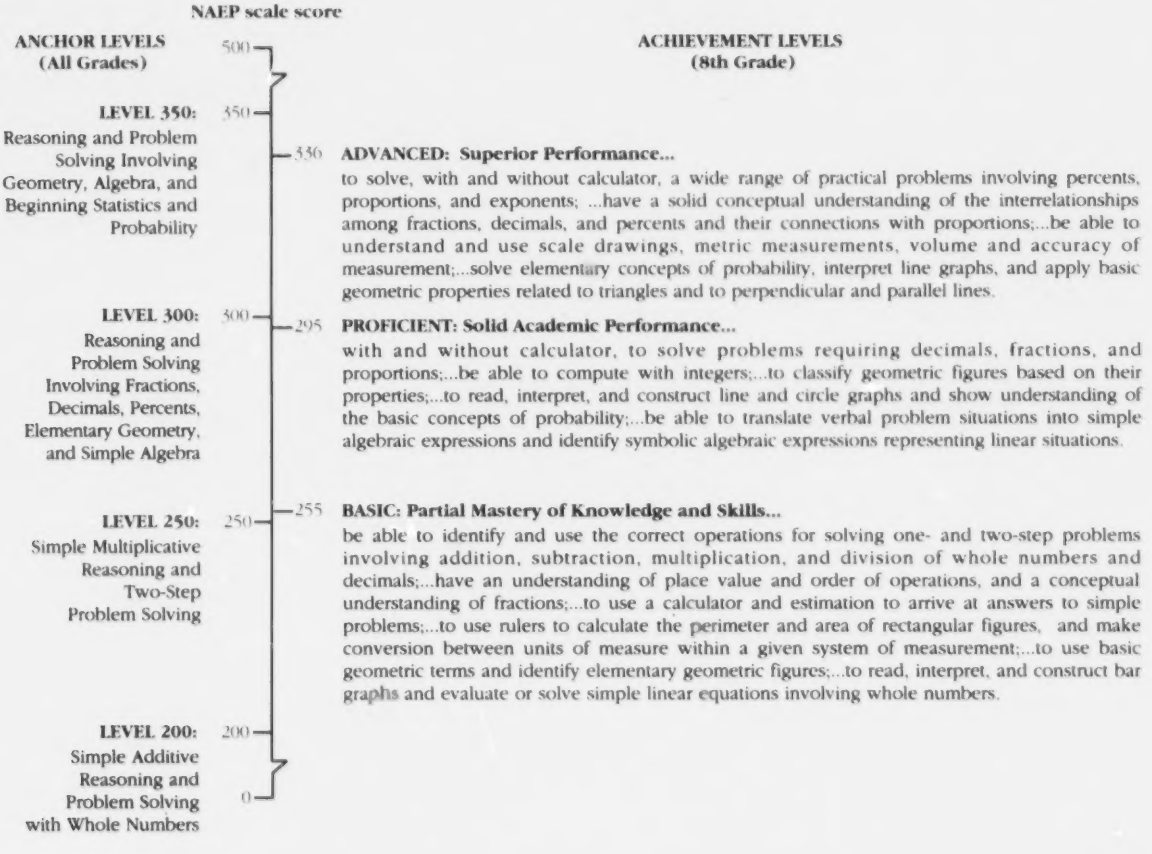
How Useful Are the Achievement Levels for Reporting Educational Progress?

The Panel is investigating the usefulness of the achievement levels for reporting as part of its ongoing study of the impact of NAEP reports, and plans to have at least a partial answer to this question in the spring, 1992. Panel researchers are in the process of asking a sample of national, state, and local policymakers and educators to judge the utility of the achievement levels in comparison to mean NAEP scale scores and anchor levels in reporting the results of the 1990 assessment.

In the meantime, the Panel placed the achievement levels and the anchor levels on the NAEP scale, thereby providing a side-by-side comparison of the two sets of levels. When examined in this light, as in Figure 3, the achievement levels do not offer appreciably more information than do the anchor levels. As the figure makes clear, the new achievement levels are scarcely more than an additional, albeit new, set of cutpoints on the NAEP scale.

Figure 3

A comparison of the 1990 8th grade achievement levels with the NAEP anchor levels.


NAEP scale score
ANCHOR LEVELS (All Grades)
500
LEVEL 350: Reasoning and Problem Solving Involving Geometry, Algebra, and Beginning Statistics and Probability
350
LEVEL 300: Reasoning and Problem Solving Involving Fractions, Decimals, Percents, Elementary Geometry, and Simple Algebra
300
LEVEL 250: Simple Multiplicative Reasoning and Two-Step Problem Solving
250
LEVEL 200: Simple Additive Reasoning and Problem Solving with Whole Numbers
200
0
ACHIEVEMENT LEVELS (8th Grade)
356
ADVANCED: Superior Performance...
to solve, with and without calculator, a wide range of practical problems involving percents, proportions, and exponents; ...have a solid conceptual understanding of the interrelationships among fractions, decimals, and percents and their connections with proportions;...be able to understand and use scale drawings, metric measurements, volume and accuracy of measurement;...solve elementary concepts of probability, interpret line graphs, and apply basic geometric properties related to triangles and to perpendicular and parallel lines.
295
PROFICIENT: Solid Academic Performance...
with and without calculator, to solve problems requiring decimals, fractions, and proportions;...be able to compute with integers;...to classify geometric figures based on their properties;...to read, interpret, and construct line and circle graphs and show understanding of the basic concepts of probability;...be able to translate verbal problem situations into simple algebraic expressions and identify symbolic algebraic expressions representing linear situations.
255
BASIC: Partial Mastery of Knowledge and Skills...
be able to identify and use the correct operations for solving one- and two-step problems involving addition, subtraction, multiplication, and division of whole numbers and decimals;...have an understanding of place value and order of operations, and a conceptual understanding of fractions;...to use a calculator and estimation to arrive at answers to simple problems;...to use rulers to calculate the perimeter and area of rectangular figures, and make conversion between units of measure within a given system of measurement;...to use basic geometric terms and identify elementary geometric figures;...to read, interpret, and construct bar graphs and evaluate or solve simple linear equations involving whole numbers.

Data from the 1990 TSA provide further evidence that the new achievement levels provide little new information. In Figure 4, the percentage of children who scored at the basic level or higher is plotted against the percentage of children who scored above the 250 anchor level for the 34 states that agreed to have the achievement levels reported. As would be expected due to the close proximity of the 250 anchor level to the grade eight basic achievement level of 255, the plot of points falls nearly on a straight line, verifying that the results of the two methods of reporting are virtually identical. The same conclusion is apparent in the results shown in Figure 5 where the percentage of children who scored at the proficient level or above is plotted against the percentage of children who scored above the 300 anchor level.

Figure 4

For each state, percent of students scoring at the basic level or above versus the percent scoring at anchor level 250 or above

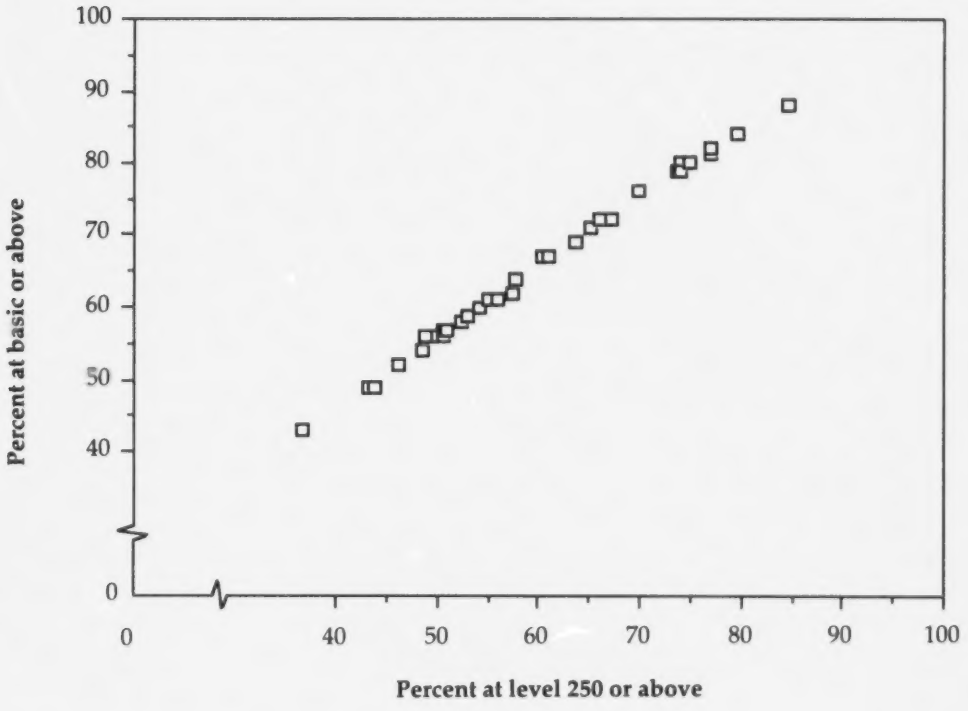
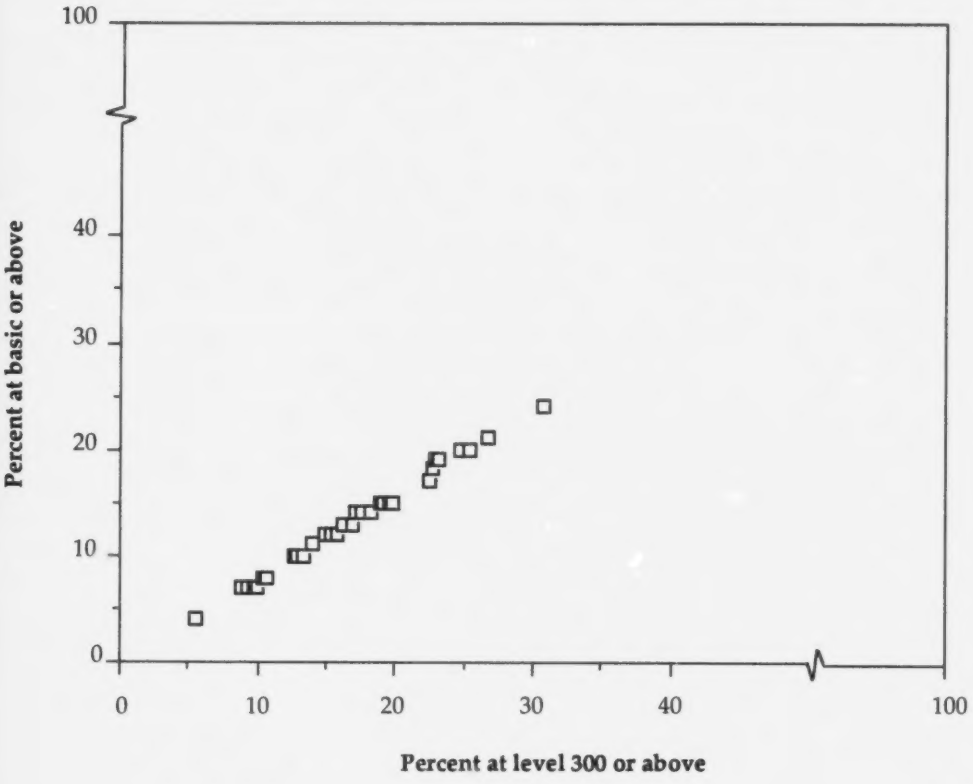


Figure 5

For each state, percent of students scoring at the proficient level or above versus the percent scoring at anchor level 300 or above



All three figures suggest that the new basic, proficient, and advanced cutting points provide little, if any, additional meaning beyond that established by the anchor levels definitions. NAEP has provided trend lines for over 20 years. If NAEP is to continue to provide consistent and useful information about how our students are doing, one can question whether it is useful to change the focus from actual achievement to aspirations. The Panel is in the process of gathering information from policymakers and educators at the national, state, and local levels to inform a recommendation about the usefulness of the achievement levels as a reporting device. The Panel will report the results in its evaluation of the 1992 TSA. In the meantime, the Panel strongly suggests that NCES report the results from the 1992 assessment using the NAEP scale score and its anchor levels, whether or not the 1992 achievement levels are used.

While the Panel understands the attractiveness of a single, global mathematics score, substantial concern exists in the mathematics education community about the wisdom of reporting NAEP mathematics results using the global scale.[13] Much of the current

[13] E.A. Silver, P.A. Kenney, and L. Salmon-Cox, "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford CA: National Academy of Education). 1992.

reform movement in mathematics is based on the assumption that there are multiple aspects to mathematics achievement. As a result, educators report finding the NAEP subscores more useful than the global NAEP score. Because NAGB has recommended that, beginning in 1992, the achievement levels are to be the primary vehicle for reporting NAEP results, the subscores and their value to educators are in danger of being lost—a serious loss in the Panel's view. The Panel strongly suggests that the results of the 1992 assessment be reported by subscore as well as by the total NAEP scale score.

The National Education Goals Panel's Decision to Report Competency Levels Instead of Achievement Levels

NEGP reported the NAEP results on September 30, 1991, in yet another form: as "competency levels."[14] "Competent" was defined by NEGP as the combination of NAGB's proficient and advanced achievement levels and "not competent" as the combination of the basic and below basic levels.[15] The Panel believes that little was gained by collapsing the data in this way; indeed, it increases the likelihood for confusion, which could arise in two ways. First, it might not be clear to all readers that the achievement levels and the competency levels are based on the same set of data. Second, what NEGP defines as "not competent" includes NAGB's basic level, a level defined as "partial mastery of knowledge and skills." If a student has mastered *some* of the material in the assessment, by this definition the student has demonstrated some degree of competence. Because of the potential for confusion, the Panel suggests that NCES, NAGB, and NEGP agree on a single format for releasing the achievement levels data in future years.

The Need for Further Research and Study on the Setting of Educational Standards and Achievement Levels

NAGB's standards-setting effort was unique—it was the first-ever attempt at setting achievement levels for the nation. The Panel understands NAGB's attempt to leverage needed change in our schools through the use of the achievement levels. Having said this, the Panel has deep concerns about the results of that effort.

During the summer of 1991, NAGB issued a Request for Proposals (RFP) to set achievement levels for the 1992 assessments in mathematics, reading, and writing, and awarded the contract in late September. Although the process is essentially the same as the process used in 1990, several worthwhile changes for the 1992 achievement-levels-setting process have been suggested by the contractor.[16] First, the contractor proposes to overlap the panels of judges in order to provide cross-grade, cross-content area linkages. Second, the "panelists will take one form of their grade-subject

[14] A figure showing the dates on which the various reports using the 1990 national and state NAEP data is shown in Chapter 5, p. 45.

[15] National Education Goals Panel, *The National Education Goals Report: Building a Nation of Learners* (Washington, D.C.: 1991), 220.

[16] American College Testing Program, *Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading and Writing* (Iowa City, IA: 1991), 6-35.

area assessment under actual NAEP testing conditions, and will receive a personal, confidential report of their performance," thereby ensuring that the judges have first-hand knowledge of the assessment and the type of items that comprise it. Third, the contractor intends to provide feedback to each panelist in an attempt to increase the internal consistency of each judge's ratings. Fourth, the evaluation of the entire process will be undertaken by "an internal quality control team and review panel of measurement experts...that will continually monitor the design and implementation of the levels-setting process."[17]

In spite of these positive changes, the Panel is concerned that the proposal for setting the 1992 achievement levels may continue a process of unknown utility and validity. A number of elements in the proposed procedure foster this concern. First, as was true in 1990, the items for the 1992 assessment will have been developed prior to setting the 1992 achievement levels. As a result, there is the danger that the item pool will not adequately support the achievement levels set, a problem that occurred at the grade 12 advanced level in 1990. As noted earlier in this report, there should be a close relationship between the frameworks, the items, and the use of the items. Because the items for the 1992 NAEP have already been determined, the setting of the 1992 achievement levels perforce *cannot* be closely tied to the selection of the 1992 items. Second, the RFP did not require the contractor to address the issue of whether the 1992 levels should be linked to the 1990 levels, nor is the contractor required to show how the design can be used to link the 1992 results to achievement levels set in future years. Without this linkage, there is no way to guarantee that the levels set from assessment to assessment can be meaningfully compared. ***Given the demonstrated value of NAEP trend data, the Panel strongly emphasizes the importance of ensuring that results of one assessment will be comparable with those from future assessments.*** Finally, the Panel is concerned that the RFP did not call for an external evaluation of the entire 1992 standard-setting process. Given the problems that occurred with the 1990 assessment and the great importance of the achievement-levels-setting process, the procedures for setting achievement levels should be evaluated by a distinguished team of external evaluators as well as by an internal team.

These issues are mostly technical concerns, all of which NAGB and its contractor could resolve. However, there is a larger, more important issue which requires investigation before the achievement levels are used on a continuing basis.

The Need to Integrate Setting Achievement Levels with Other Standard-Setting Activities in the Nation

As mentioned earlier, the National Council on Education Standards and Testing (NCEST) has recommended the establishment of national content and performance standards. Because NCEST went out of existence with the completion of its report to Congress, NCEST recommended that Congress establish a new council on national standards and assessment to coordinate the development of new education standards and then to certify that they are "world class." Assuming that Congress enacts NCEST's recommendation, it is extremely important that NAGB ensure that the development of the achievement levels is done in close coordination with NCEST's successor as it seeks to develop national standards and assessments. Finally, the Panel

[17] *Ibid.*, 25.

again stresses the need for studies that validate the descriptors NAGB uses to illustrate its achievement levels.

> The use of achievement levels in reporting NAEP results can be of value both to the nation and to the states participating in the TSA, by providing succinct, accessible information for public discussion of education and reform. However, it is essential that NAGB's future efforts at setting achievement levels be closely coordinated with the reports and recommendations of the council proposed to succeed NCEST. This will be essential to ensure that a coordinated set of educational standards informs policy and discussion. The Panel further recommends that when setting achievement levels for future assessments, NAGB conduct studies to validate the descriptors used to illustrate the achievement levels—especially those implying outcomes of future performances, such as success in postsecondary education and employment.

5 Reporting the Results to the Public and the Press

Overview

In this chapter, the Panel examines the reporting of the 1990 trial state mathematics assessment. The reports and other materials released by the National Center for Education Statistics (NCES) in June 1991 are described first. Special attention is paid to the manner in which the state-by-state comparisons were presented, because all parties were concerned that simplistic state comparisons could be misleading and, by focussing on inappropriate criteria, ultimately harmful to education.

The ways in which the press reported the results are also examined. Both the extent and the accuracy of the reporting are considered in the Panel's analyses—the latter information on accuracy is helpful particularly in evaluating the extent to which the materials and press conference prepared by NCES were successful in interpreting the statistically complex results of the assessment for a lay audience.

A brief review of the second, September 1991, presentation of the assessment results, in which the National Assessment Governing Board (NAGB) achievement levels were used for the first time, is also provided. The chapter concludes with an examination of how the results have been used by the states during the first few months following the release.

Based on its examination of the reporting process for the 1990 Trial State Assessment (TSA), the Panel concludes:

- *NCES and its contractor, the Educational Testing Service (ETS), did an excellent job of presenting the results of the 1990 TSA to the public in June 1991. The assessment yielded an enormous wealth of data that challenges the reader both in terms of its sheer volume and technical complexity. The decision to release multiple reports, each at a different level of summarization, helped enormously to keep the data manageable.*

- *NCES and ETS are also to be commended for their work in presenting the state-level data in accurate and fair ways, while at the same time developing reporting formats that discouraged simplistic conclusions about state comparisons.*

- *In the context of this generally fine job of reporting, the Panel notes that NCES engendered criticism from a number of state representatives when it reported grade-level equivalents for NAEP scale anchor levels without previously sharing its intentions to use the grade-level equivalents with the states.*

- *By and large, the press reported accurately the central message of the June 1991 data release—that mathematics performance in the United States is inadequate and needs to be improved. A high proportion of the published articles, however, discussed the state rankings without providing the statistical information necessary to help the reader interpret results for states with overlapping performance. In addition, much of the wealth of potentially useful data about the teaching and learning of mathematics was lost in the media reporting.*

- *A second release of data from the 1990 TSA occurred in late September 1991 when NAGB, NCES, and the National Education Goals Panel (NEGP) all released reports containing TSA or other NAEP data within a four-day period. Subtle differences in the ways in which the data were reported by the different agencies exacerbated the difficulties inherent in assimilating such a large volume of educational data.*

- *The June 1991 release of results provoked discussions in state departments of education, state boards of education, and state-level education organizations. State mathematics specialists and testing directors interviewed in late summer reported that the results were even then providing support for reforms contemplated or already underway, and indicated that the impact of the state assessments could be expected to increase once successive rounds of testing had occurred and the states could review their performances over time. The Panel is continuing to investigate the impact of the September release and the perceived utility of the different reporting formats used in June and September.*

The Timetable for the Release

Although the data collection for both state and national NAEP was completed in spring of 1990, the data were not ready for public release until June 1991. This somewhat lengthy delay was occasioned largely by the complexities of analyzing and formatting information on the first Trial State Assessment (TSA). The reporting cycle for the 1992 TSA is scheduled to be considerably shorter. Figure 6 provides an overview of the NAEP results that were released during 1991; although some of these results are not directly related to the TSA, they are relevant because they were released concurrently with the TSA results and strongly influenced the way in which the TSA results were reported in the press.

Figure 6

Release of 1990 NAEP results

Date	Agency	NAEP Results
June 6, 1991	National Center for Education Statistics (NCES)	Main 1990 mathematics assessment—state and national results (scale scores and anchor levels)
September 27, 1991	NCES	National trend data through 1990—mathematics, reading, writing, and science (scale scores and anchor levels)
September 30, 1991	National Assessment Governing Board (NAGB)	Main 1990 mathematics assessment—state and national results (three achievement levels)*
September 30, 1991	National Education Goals Panel (NEGP)	National trend data and 1990 achievement level results for the nation and the states (three NAGB achievement levels collapsed to one)

* Included data only for states agreeing to have their data re-analyzed and released using the achievement levels; three states, Arizona, Illinois, and Montana, chose not to participate.

As shown in the figure, state and national data from the main mathematics assessment were made public by the National Center for Education Statistics (NCES) on June 6, 1991. The traditional NAEP scale scores and anchor levels were used to describe the data. Driven by the schedule for the first *National Education Goals Report*, a second major data release occurred in late September. At that time, the National Assessment Governing Board (NAGB) presented the 1990 national and TSA results using its new achievement levels. NCES presented national NAEP *trend* data for the core disciplines of mathematics, reading, writing, and science.[1] And the National Education Goals

[1] One of the great strengths of the NAEP program has been its ability to present *trend* data on the achievement of U.S. students. This requires NAEP to test cohorts of students in successive years using overlapping pools of items so that results from different years can be compared with confidence. In preparing for the 1990 assessment, the decision was made to *break* the existing trend line in order to accommodate more radical changes in test content and format, consonant with changing concepts of mathematics education and assessment. NCES, however, continues to test smaller "bridge" samples using the older item pool so that it can maintain the trend data. Eventually, when the new trend line is firmly in place, these bridge samples will be discontinued.

Panel (NEGP) presented both the trend data and the state and national mathematics results. For the latter, NEGP used its own variant of the NAGB achievement levels, in which it collapsed the three NAGB levels—advanced, proficient, and basic—by combining advanced and proficient into one level designated as competent, and combining basic and below basic into a single level designated as not competent.

The Results Released in June

The press materials developed by NCES and its contractor, the Educational Testing Service (ETS), were intended to present the richness of the 1990 national and state results on mathematics achievement to the public, while at the same time averting erroneous conclusions about differences in performance based on simplistic rankings of the states. The press materials also included commentary by Secretary of Education Lamar Alexander and members of NAGB, who portrayed the results as falling short of the levels of mathematics achievement required in today's world of high technology.

The 1990 NAEP results for all participating states and the nation were reported in *The State of Mathematics Achievement*—a volume of more than 500 pages, summarized in an executive summary of 41 pages, and presented in a press conference that included statements from NCES and ETS staff, members of NAGB, and Secretary Alexander. Separate volumes focussing on the results for each participating state were also released.

The extent of the data is best portrayed by a brief review of the contents of the composite volume. This volume was divided into two major sections: one addressing the national results for grades 4, 8, and 12 and the second describing the TSA results for grade eight. The section on national results was organized under these general headings:

- Overall mathematics proficiency for the nation
- Overall mathematics proficiency results for population subgroups
- Mathematics proficiency by content area for the nation and population subgroups
- Course-taking patterns in the mathematics content areas for the nation and population subgroups
- Student performance on constructed-response questions
- Instructional approaches
- Calculators and computers
- Instructional time and emphasis
- Students' perceptions of mathematics
- Characteristics of mathematics teachers

In the second section, similar information was reported for the grade eight TSA, and an additional chapter reported state-level proficiency results by measures related to academic emphasis in the home—including hours of homework, television watching, school attendance, and parents in the home. The results for the 37 states, two territories, and the District of Columbia were presented state by state and in the context of national and regional results.

Overall, the composite volume included 199 tables and 22 figures plus 4 appendices with information on state contextual background factors, including co-statistics from external sources, state-level participation rates, descriptions of the assessment procedures and the scale-anchoring process, and example items for each anchor level. The executive summary and the press release prepared by Emerson Elliott, acting commissioner of education statistics, and Gary Phillips, acting associate commissioner, education assessment, offered selected findings from each of the main topics covered in the composite report. Separate statements by NAGB board members and Secretary Alexander were intended to highlight policy implications and provide thoughtful commentary on the results.

In reviewing these volumes and related press materials, the Panel concludes that NCES did an excellent job with a difficult task. The decision to release multiple reports, each at a different level of summarization, helped enormously to keep the data manageable. Tables and graphs, although somewhat cumbersome due to the density of the data, nevertheless did a good job in focussing the reader on the appropriate and interesting comparisons between and within states.

The success of this first ever release of state-level NAEP data was due, at least in part, to efforts to keep the states closely involved in the reporting process. NCES and ETS worked closely with the states to identify the most useful reporting formats, distributing draft versions of the reports (with mock data) for early review, and providing multiple opportunities for feedback through group meetings as well as more informal communications. NCES, working with the Council of Chief State School Officers (CCSSO), also provided advance interpretive materials to the states and hosted a pre-release briefing for state education representatives on June 3, 1991. At that time, the representatives had the opportunity to review all of the final NCES reports and press statements and to prepare for simultaneous data releases in their home states.

State-by-State Comparisons

Another challenge faced by NCES was to formulate reporting mechanisms that would foster appropriate and responsible comparisons between the states. As the National Academy of Education noted in its 1987 commentary on the Alexander-James report that laid the groundwork for the current NAEP state assessment, "Many factors influence the relative rankings of states, districts, and schools. Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts."[2] In the 1990 mathematics assessment, differences in average proficiency between many of the states were so small as to be of no practical or statistical significance. Moreover, many aspects of state performance are not well captured by a single summary

[2] National Academy of Education, "Review of the Report by a Committee of the National Academy of Education," in *The Nation's Report Card: Improving the Assessment of Student Achievement* (Washington, D.C.: 1987), 59.

measure such as overall average proficiency. State comparisons for relevant subpopulations such as racial-ethnic groups or advantaged/disadvantaged communities are other important measures of differential state performance, and they can even, on occasion, produce results that seem to contradict overall averages. For example, Texas had an average proficiency for all students that was two points lower, albeit not significantly different, than Rhode Island. However, for each of the three racial-ethnic groups with sufficient sample size to permit reliable estimates, Texas students had a *higher* average proficiency than their Rhode Island counterparts—by 8 points for white students, by 9 for black students, and by 17 for Hispanic students.[3]

Another perspective on the differential performance of racial-ethnic groups across states is offered in Tables 5 and 6, where state rankings based on the performance of their white students is contrasted to state rankings based on the performance of their black or Hispanic students. As can be seen in the first of these tables, only two of the seven states that rank in the top quartile based on the performance of their black students also rank in the top quartile based on the performance of their white students. One state (Kentucky) in the top quartile for black student performance is actually in the *lowest* quartile based on white student performance. As shown in Table 6, the shifts in state ranking based on Hispanic student performance are also quite dramatic in some instances. For example, two of the states that ranked in the lowest quartile based on white achievement were in the third quartile based on Hispanic achievement.

Table 5

Comparison of state standing based on black and white student achievement

	State standing based on black student achievement			
State standing based on white student achievement	Lowest quarter	Second quarter	Third quarter	Highest quarter
Highest quarter			5	2
Third quarter		2	2	3
Second quarter	2	4		1
Lowest quarter	5	1		1

Table based on 28 states with sufficient numbers of black and white students to permit the calculation of separate statistics for each group.

From R.L. Linn, L. Shepard, and E. Hartka, "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns," *in Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

[3] This seeming paradox results from the differences in the racial-ethnic compositions of the two states, combined with overall differences in performance between racial-ethnic groups. Thus, Hispanic and black students in Texas may have scored higher than their Rhode Island counterparts, but they still scored lower than white students in either Texas or Rhode Island. Since the lower scoring racial-ethnic groups comprise a higher proportion of the Texas population, the state as a whole scored lower.

Table 6

Comparison of state standing based on Hispanic and white student achievement

	State standing based on Hispanic student achievement			
State standing based on white student achievement	Lowest quarter	Second quarter	Third quarter	Highest quarter
Highest quarter		2	1	6
Third quarter	1	2	3	3
Second quarter	1	5	3	
Lowest quarter	8		2	

Table based on 37 participating states--does not include District of Columbia, Guam, or the Virgin Islands.

From R.L. Linn, L. Shepard, and E. Hartka, "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns."

Just as overall state means do not necessarily provide accurate information about the performance of particular subgroups within the states, they do not provide very much information about the overall *distribution* of performance of students within the states. They do not tell us, for example, that the top 10 percent of students in Hawaii score 95 points higher than the bottom 10 percent, while the corresponding difference for Montana is only 66 points.

The Panel's study indicates that NCES and ETS did a good job in developing formats that discouraged simplistic conclusions about state comparisons and reported the state-level data in accurate and fair ways. The results for the 37 states, two territories, and the District of Columbia were presented state by state and in the context of national and regional results. Subpopulation results for each state were presented prominently to facilitate comparisons among states with different mixtures of ethnic and socioeconomic groups. State-level results were also reported in terms of the percentages of students scoring at or above each of the four anchor levels and in terms of the scores achieved by students near the top and bottom of each state's distribution. In addition, standard errors were reported along with mean levels of achievement, and statements to the press emphasized repeatedly that it would be misleading to rank-order the states and territories from first to 40th because the rank-orderings for many states were not statistically different from one another. In interviews conducted by the Panel, 80 percent of the state assessment and mathematics specialists commented favorably on the manner in which NCES handled the comparisons issue.[4]

[4] F. Stancavage, E. Roeber, and G. Bohrnstedt, "A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

The Panel also commends NCES for its efforts to provide the press with responsible graphic interpretations of the inevitable state-by-state comparisons, rather than side-stepping the issue by simply presenting the state data in random or alphabetical order and admonishing the press not to rank-order them. The primary graphic that NCES put together for presenting the state rankings was very creative and had high potential for reducing errors and confusion. However, the strength of the chart as a statistical reporting tool was its weakness as a newsmedia tool. The press response to the 'pantyhose chart'—as it became known—was that the format was too complex and unfamiliar to be reported and explained in a typical news article, even an article in a weekly newsmagazine.

Grade Level Equivalents for NAEP Scale Anchor Levels

The Panel noted that NCES created a certain level of confusion and engendered criticism with its decision to use grade level equivalents. As of the June 6, 1991, release date, the work on the NAGB achievement levels had not yet been completed. Instead, state and national results were reported in terms of the more traditional NAEP scale with anchor levels at 200, 250, 300, and 350—in use since the early 1980s. In an effort to provide familiar descriptors for these NAEP anchor levels, the contractor and NCES decided to associate each anchor point with a grade-level referent as well as to provide simple prose descriptions of the mathematical tasks mastered at each level. A panel of 19 mathematics educators was assembled some months prior to the release of the data to undertake this task, and their conclusions were summarized in the *State of Mathematics Achievement*, as follows: "The panel...characterized Level 200 as material typically covered by the third grade, Level 250 as material generally covered by the fifth grade, Level 300 material as content introduced by the seventh grade, and Level 350 as content generally covered in high school mathematics courses in preparation for the study of advanced mathematics."[5] The report then presented data on the percentages of 4th, 8th, and 12th graders that scored at each of the various anchor levels.

The approach taken to generating these grade-level referents was not well considered. No evidence is presented that those teachers who assigned the ratings constituted a sample whose combined curricular experience was representative across either grade levels or geographical regions. In fact, it is hardly likely that a sample as small as 19 could reliably represent either dimension. In addition, judgments as to the grade levels at which topics are commonly *first covered or introduced* should not be confused with judgments about the grade levels at which educators would expect *mastery*. It is common for teachers to present material at more than one grade level before students are expected to master it.

State representatives meeting at the June 3, 1991, pre-release briefing were surprised to find the NAEP scale-score anchor points (and hence state performance) prominently described in terms of these grade-level equivalents. NAEP had never before been reported using grade-level equivalents, yet NCES had failed to review these equivalents with the participating states or CCSSO in advance. This unfortunate lapse occurred in spite of the fact that, as noted above, there had been extensive efforts made to keep the states informed of the plans for the reporting process and to solicit their input into these plans. NCES responded to the resultant criticisms by

[5] U.S. Department of Education, National Center for Education Statistics, I.V.S. Mullis et al., *The State of Mathematics Achievement* (Washington, D.C.: 1991), 6-7.

removing references to grade levels from the more prominent graphics used at the press conference on June 6, 1991, but was unable to alter the printed reports. Moreover, the press statements of Secretary Alexander and NAGB member Francie Alexander referred to the mathematical tasks that characterize Level 300 as "something eighth graders should know and be able to do" or "material commonly introduced by the seventh grade." Consequently, the press coverage following the conference contained numerous references to direct grade level equivalents; for example, the *Houston Chronicle* ran the headline: "A minus sign on math skills: Most Texas eighth graders test below seventh grade."

In interviews conducted later in the summer, several state assessment and mathematics specialists indicated that the inclusion of the grade-level equivalents had created a reporting problem for them because the grade-level equivalents had never been reviewed with them in advance, were not supported by technical reports explaining their development, and could not be aligned with curricular practices in their states.

What the Press Reported

Included in the materials distributed at the June 6, 1991, press conference were press releases designed to sound an alarm about mathematics achievement in America. Secretary of Education Lamar Alexander stated: "The big news is: none of the states are cutting it." The NAGB press release included several opinions, beginning with a statement by Richard A. Boyd, chairman, pointing out that the results "do not portray a happy picture....let's use today's results to examine where our deficiencies may lie and then get to work to correct them." In another statement Francie Alexander, then a member of NAGB, said, "Unfortunately, the data show we have an enormous way to go even in the states which appear to be doing *relatively* better than others." Carl J. Moser, another NAGB member, focussed on the way the factors of home and school intersect and said, "What we need is a national effort to promote learning and to value it....It means a change in our values and expectations." The common theme was to motivate the American public to focus their attention and efforts on improving mathematics achievement.

This theme was picked up and reinforced by others in the education policy community. For example, CCSSO headlined its release with "National Mathematics Campaign Imperative," and quoted Gordon M. Ambach, CCSSO's executive director: "The results from the 1990 math NAEP indicate a tough, uphill journey for students...a national campaign is imperative."

By and large, the press reported accurately the central message portrayed by NAGB and Secretary Alexander (as well as many of the states)--that mathematics performance is inadequate and needs to be improved. Typical headlines included "U.S. Students Weak in Math, Study Reveals: Most computer-age kids know only basic arithmetic" (*San Francisco Chronicle*), "Report charts U.S. students' shortcomings in math" (*Philadelphia Inquirer*), and "Math's a problem, tests show; 'we compare favorably,' but educators aren't happy" (*Rocky Mountain News*, Denver). When questioned, nearly two thirds of the state assessment and mathematics specialists indicated that the news media accurately conveyed the underlying thrust of the assessment findings, and 40 percent said that the news media did a good job of presenting the information.

About a quarter of the respondents expressed dismay at the lack of response shown by the news media in their states, calling the release a media "non-event."

State Comparisons

Four of the most influential newspapers in the nation, the *New York Times, The Wall Street Journal, The Washington Post,* and *USA Today,* rank-ordered the states in their news articles despite caveats from both NCES and Secretary Alexander not to do so. The *New York Times* warned against "using the data as a simple ranking because of the many variables involved in the calculations and the overlapping performance of many states," while the *Post* mentioned that many of the tabled differences were not statistically significant. Neither paper, however, provided any further guidance as to *which* of the differences were significant or not significant. Moreover, the *USA Today* article contained the incorrect interpretation that "Eighth graders in North Dakota, Montana, Iowa and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, D.C." Although the mean scale score of top-scoring North Dakota's eighth graders *is* 21.6 percent higher than the mean scale score of the District of Columbia's eighth graders, for the *USA Today* interpretation to be correct would require that a NAEP scale score of zero represent total absence of knowledge and that the NAEP test measure the whole of subject area knowledge.

In a review of 55 articles published in the top 50 newspapers (by circulation) during the week the results were released, the Panel found that state rankings were mentioned in about two thirds of the articles. Ten accounts presented the average scores for all states in table form, either ordered by state rank or with rank indicated, and two others presented partial or complete lists of the states ranked by the percentage of students scoring at Level 300. Fewer than half of these articles—and only three of those that used tabular displays—provided appropriate interpretation of these statistics.

By comparison, when describing the ranking of a *single state,* at least half the reporters were able to formulate appropriate interpretations based on the language provided by NCES and their own state departments of education. For example, the *Hartford Courant* reported that "Allowing for possible statistical errors, researchers said that only five states scored significantly higher than Connecticut and 12 others had roughly equal scores," and the *St. Petersburg Times* (Florida) wrote that "Nineteen states scored higher than Florida, sixteen made similar marks and four ... posted significantly worse scores." By contrast, however, The *Miami Herald Tribune* reported the Florida data as follows: "37 states, 2 territories and the District of Columbia participated and were ranked in 3 groups, with 19 in the first group, 17, including Florida, in the second, and 4 in the third." These "groups" were not part of the analysis or reporting design, as implied by the article, but simply the numbers of states who ranked higher, lower, or the same in relation to Florida.

Thus, it appears that efforts to help the press articulate statistical concepts can be beneficial if the interpretations are offered at a level the press can communicate to its readers. The Panel recommends that NCES continue to provide the media with specific examples of interpretations that are (or are not) supported by the results as well as continue efforts to construct new tables and graphs that effectively display the data. The Panel also

recommends that newly developed reporting devices be presented for comment to the press or media specialists in a series of focus groups, prior to the release of future reports. These efforts may enable NCES to determine the types of data and displays that best satisfy the reporting needs of the media while preserving the accuracy of the data.

Press Coverage and the NCES Release Schedule

As noted above, *The State of Mathematics Achievement* contains an overwhelming amount of data, and even the highly condensed press release presented highlights of more than a dozen separate aspects of national and state performance. Yet news doesn't stay news for long, and most newspapers published only a single article on the results. It is not surprising, therefore, that the news reports generally failed to capture the full richness of the material. Most of the articles focussed on the overall results for the nation and/or the paper's home state, but only a few provided detailed discussions of performance differences for subpopulations within their state or between their state and others. Similarly, about half the articles touched on the effects of TV watching or the amount of time spent on homework, and about a quarter reported briefly on the positive effects of teacher specialization and parent educational levels. However, significant coverage of some of the most interesting and potentially useful aspects of these contextual data, such as the relationship of student performance to differences in mathematics curriculum or teacher preparation, was scant at best.

In light of these observations, the Panel concludes that releasing all of the data at one time may be a mistake. The Panel suggests that NCES consider spreading the release of the data over several months. After the first report on the state-by-state results, individual reports could be released concentrating on the relationship of outcomes with such variables as course-taking patterns, instructional time devoted to particular subjects, teacher experience and preparation, and school characteristics. It would be important, however, to insure that the initial report be sufficiently inclusive to provide a fair context for the comparisons between the states. States with large minority populations, for example, would probably feel strongly that information on the differential performance of various racial or ethnic subgroups should be included in the initial report.

The Achievement Levels and the National Goals

At the end of September, the public and the press were again presented with a wealth of data based on state NAEP, national NAEP, and other educational indicators. As in June, because the data were all reported at the same time, the press tended to cover them in a single news story based on a composite of the findings.

The data release began on Friday, September 27, 1991, when NCES presented up to 20 years of national NAEP trend data for mathematics, science, reading, and writing. (As noted above, trend data are not yet available for the states, because 1990 constituted the first-ever state-level assessment of this kind.) On Monday morning, September 30,

1991, the NEGP presented their first report card for the nation and the states. Student performance data from the 1990 TSA—recalculated in terms of the NAGB achievement levels that define what students *should* know at grades 4, 8, and 12—were featured prominently as national and state indicators of Goal Three: Student Achievement and Citizenship in *America 2000*. The national NAEP trend data (which use the traditional, behaviorally anchored NAEP scale and anchor levels) and the 1990 NAEP/TSA data on mathematics instructional practices were also included as "additional important information" relevant to Goals Three and Four. (Goal Four states that by the year 2000, U.S. students will be first in the world in science and mathematics achievement.) These data were in addition to many other, non-NAEP indicators that were presented to measure progress towards each of the six goals put forward by the President and governors.

Finally, later in the day on the 30th, NAGB held its own press conference to elaborate on the achievement levels. Four categories based on the three achievement levels were presented for each grade level: below basic, basic, proficient, and advanced. By contrast, the National Education Goals Panel had collapsed these into two categories—not competent and competent—for their main presentation, although the original NAGB achievement levels were used in the "additional important information" section of their report. NAGB also presented the achievement-level results broken down by sex, race or ethnic group, type of community, parents' education, public versus private schools, and, for grade 12, semesters of high school mathematics.[6] The printed report (available at the press conference) included a companion volume of national and state results on released items, showing the percentages of students at each achievement level who answered the items correctly.

A primary theme for all speakers was that the nation's students are doing about as well as they did 20 years ago, but that this is not good enough for today's high-tech world. Hence the decision was made to set high competency standards (attained by fewer than one in five in 1990) towards which students should strive. Discussion of the controversy that has surrounded the technical aspects of the achievement-levels-setting process was minimized by NAGB, which noted that the standards-setting process will be continued into additional subject areas with the 1992 assessment, and that "[the Board] will also review these [1990] mathematics levels carefully, because this effort is a trial, and possibly make some adjustments before using them to report the 1992 NAEP mathematics assessment."[7]

As with the state comparisons based on the average proficiency scores or the NAEP anchor levels released in June, caution should be exercised in interpreting the rankings based on achievement levels. For example, as shown in Table 7, a number of states appear to be doing significantly better if compared on the percentages of students who achieve the basic level than if compared on the percentages achieving the proficient level, while others would be favored by rankings based on the percentages of students achieving at the advanced level. Oklahoma, for example, ranks 16th if rated on the basic level but 24th when rated on the proficient level, while Virginia, which ranks 13th on the proficient level, ranks first if rated on the advanced level.[8]

[6] The latter two breakdowns apply to the national data only, since neither private school students nor twelfth graders were tested in the 1990 TSA.

[7] National Assessment Governing Board, Richard A. Boyd, *Statement on the Levels of Mathematics Achievement*, press release (September 30, 1991).

[8] Ranks based on the 34 states that chose to have their 1990 mathematics results reanalyzed in terms of achievement levels.

The size of the sampling error and the extent of overlapping performance are also important to consider, particularly when one examines achievement at the "advanced" level. Because the percentages of students achieving at this level are very small relative to the size of the sampling errors, there are *extensive* overlaps in state performance. To understand the extent of this overlap, one needs to construct confidence intervals around the estimated percentages for each state, using the information shown in Table 7.[9] For example, the confidence interval for Virginia (the top ranked state based on performance at the advanced level) extends from 0.9 percent (2.3 minus 1.4) to 3.7 percent (2.3 plus 1.4). This overlaps the confidence intervals for all states down through Ohio, California, and Wyoming, which are tied for the 21st rank. By contrast, ranks based on the percentages of students achieving at or above the "basic" level are much more stable because so many more students reach this level. For example, the confidence interval for the top ranked state, North Dakota, extends from 81.8 percent (84.8 minus 3.0) to 87.8 percent (84.8 plus 3.0). This overlaps only the confidence interval for Iowa, the second ranked state.

Table 7

Percentages of students in each state performing at the basic, proficient, and advanced levels (state rankings based on each measure shown in parentheses)

State	% Basic or above	% Proficient or above	% advanced
Alabama	46.2 ± 3.0% (31)	8.8 ± 1.6% (33)	.4 ± .4% (26)
Arkansas	51.0 ± 2.6% (25)	9.7 ± 1.8% (31)	.2 ± .2% (30)
California	50.6 ± 3.6% (26)	13.9 ± 2.4% (23)	.7 ± .4% (21)
Colorado	66.1 ± 2.8% (11)	18.2 ± 2.0% (14)	.6 ± .2% (24)
Connecticut	67.3 ± 2.8% (10)	23.1 ± 2.0% (6)	2.0 ± .6% (2)
Delaware	54.2 ± 2.6% (22)	16.1 ± 1.6% (19)	1.1 ± .8% (11)
Florida	48.5 ± 3.2% (30)	12.8 ± 2.0% (26)	.4 ± .4% (26)
Georgia	52.9 ± 3.2% (23)	14.7 ± 2.8% (22)	1.0 ± .6% (15)
Hawaii	43.8 ± 2.2% (32)	12.5 ± 1.4% (27)	.9 ± .4% (16)
Idaho	73.7 ± 2.0% (8)	19.5 ± 2.2% (11)	.5 ± .6% (25)
Indiana	65.1 ± 3.6% (13)	17.5 ± 2.6% (15)	1.1 ± .6% (11)
Iowa	79.8 ± 2.4% (2)	26.7 ± 3.0% (2)	1.7 ± .8% (3)
Kentucky	50.6 ± 3.4% (26)	10.5 ± 2.2% (28)	.2 ± .2% (30)
Louisiana	36.7 ± 3.6% (34)	5.4 ± 1.6% (34)	.2 ± .2% (30)
Maryland	55.9 ± 3.8% (20)	17.1 ± 2.6% (16)	1.3 ± .6% (9)
Michigan	61.0 ± 3.0% (15)	16.8 ± 2.4% (17)	.8 ± .4% (18)
Minnesota	77.0 ± 2.4% (3)	25.4 ± 2.4% (4)	1.6 ± .6% (7)
Nebraska	77.0 ± 2.4% (3)	26.6 ± 2.6% (3)	1.7 ± .8% (5)
New Hampshire	74.0 ± 3.4% (6)	22.5 ± 2.4% (9)	1.3 ± .6% (9)

[9] Confidence intervals are constructed using information about error due to sampling. In the long run, 95 percent of such intervals will contain the true value for the population parameter of interest (in this case, the percentage achieving at the "advanced" level).

Table 7, continued

State	% Basic or above	% Proficient or above	% advanced
New Jersey	66.1 ± 3.2% (11)	22.8 ± 2.2% (7)	1.7 ± .6% (6)
New Mexico	50.0 ± 2.6% (28)	10.3 ± 1.6% (29)	.4 ± .4% (26)
New York	57.5 ± 3.8% (19)	16.2 ± 2.0% (18)	1.1 ± .8% (11)
North Carolina	43.3 ± 2.6% (33)	9.2 ± 1.4% (32)	.2 ± .2% (30)
North Dakota	84.8 ± 3.0% (1)	30.6 ± 3.8% (1)	1.8 ± 1.2% (4)
Ohio	60.4 ± 3.0% (16)	15.6 ± 2.2% (20)	.7 ± .4% (21)
Oklahoma	60.4 ± 3.6% (16)	13.2 ± 2.4% (24)	.3 ± .2% (29)
Oregon	69.9 ± 2.8% (9)	22.6 ± 2.4% (8)	1.1 ± .6% (11)
Pennsylvania	63.8 ± 4.4% (14)	19.1 ± 2.8% (12)	.9 ± .6% (16)
Rhode Island	55.2 ± 1.6% (21)	15.3 ± 1.6% (21)	.8 ± .4% (18)
Texas	52.4 ± 4.0% (24)	13.1 ± 2.0% (25)	.8 ± .4% (18)
Virginia	57.9 ± 3.2% (18)	18.9 ± 3.2% (13)	2.3 ± 1.4% (1)
West Virginia	48.9 ± 2.8% (29)	9.8 ± 1.6% (30)	.2 ± .2% (30)
Wisconsin	75.0 ± 3.2% (5)	24.6 ± 3.2% (5)	1.4 ± .8% (8)
Wyoming	74.0 ± 2.0% (6)	19.7 ± 2.0% (10)	.7 ± .6% (21)

In the long run, 95 percent of the intervals constructed as shown (e.g., 19.7 ± 2.0%) will contain the true value for the population parameter of interest. Ranks are based on the 34 states that chose to have their 1990 mathematics results reanalyzed in terms of achievement levels.

Based on data from National Assessment Governing Board, M.L. Bourque and H.H. Garrison, *The Levels of Mathematics Achievement, Initial Performance Standards for the 1990 NAEP Mathematics Assessment, Volume I: National and State Summaries* (Washington, D.C.: 1991).

Interviews with state assessment and mathematics specialists in late summer 1991 indicated that very few states anticipated that they would be making extensive commentary on the upcoming September data release. Future interviews will confirm what actually was done at that time.

How Useful Did the States Find the Reports of the Results?

The Panel commissioned a study to determine whether assessment specialists, mathematics specialists, opinion leaders, and policymakers at the national, state, and local levels found the reports useful. In reviewing these results, however, it should be understood that they are based on interviews conducted two to three months after the June data release, and *before* the release of the achievement levels. The Panel must

therefore limit its discussion to short-term impact. As discussed below, additional surveys are planned to capture more information on the impact of the reporting over time.

To ensure the usefulness of the results and reports for the states, NCES and ETS held several meetings with state-level assessment and mathematics specialists to help determine the content and formats of the reports. These meetings were very fruitful in the sense that they resulted in reports that conformed to the expressed desires of the states. Given that input, how successful was NCES in providing useful reports?

The Panel found that nearly all of the states that participated in the TSA (and several that did not) used the printed reports prepared by NCES in their state level reporting; these reports were used to inform people within the state and to assist in developing state-level press releases and press briefings. About three quarters of the participating states also availed themselves of the NCES press kit materials and closed circuit coverage of the national press conference. As useful as the materials were, however, a number of suggestions for improving them in the future were provided. For example, it was suggested that NCES might prepare a separate, brief report suitable for sharing with parents.

The results that the states found most useful were the overall level of performance on the total score, levels of performance in the separate mathematics content areas, levels of performance for various student subgroups, and levels of performance for students exposed to various instructional practices. States reported that the release of results provoked discussions particularly in the state department of education, the state board of education, and among state-level education organizations. The overall impact of the results was reported as positive by 45 percent of the respondents, mixed by 45 percent, and negative by only 3 percent, with the remainder having no opinion. Many of those who rated it as "mixed" or "negative" were primarily concerned that the results would simply be ignored by the public and the press. A few felt that the school and teacher data collected by NAEP misrepresented the level of teacher preparation in their state. Others expressed concern that the press treatment of the results would leave the public with an unduly dismal picture of their educational system.

During interviews conducted in late summer 1991, a number of the state specialists reported that the TSA results were providing strong support for reforms that were contemplated or already underway. Not surprisingly, given the brief time that had elapsed since the June release, very few reported that they had yet seen new educational reforms motivated by the TSA results or based on TSA findings, but many mentioned that the impact of the state assessments could be expected to increase once successive rounds of testing had occurred and the states could review their performance over time.

Over the coming months, the Panel will continue to investigate the longer term impact of the reporting of TSA results with further rounds of interviews, case studies, and analyses of news coverage. The next round of interviews, scheduled for January 1992, will pay particular attention to the impact and perceived utility of the various reporting formats that are now available to the public, including the achievement levels released by NAGB on September 30. This question is especially relevant in light of the current plans to use the achievement levels as the primary reporting format for 1992 and beyond.

6 The Future of State NAEP

Overview

The first of two Congressionally mandated state NAEP trials is completed. Congress has asked the Panel to develop recommendations, based on the results of that trial, about whether trial state NAEP should be continued, and if so, in what form. The Panel's recommendations are of immediate importance to Congress as it considers NAEP reauthorization and funding levels for 1994 and beyond. In this chapter, in addition to addressing the request put forth by Congress, the Panel examines whether the populations currently assessed by NAEP should be expanded to include school dropouts and a sample of young adults and whether the prohibition against reporting below the state level should be removed. Finally, the Panel explores the role state NAEP can play in the current climate of educational change and reform.

It is the Panel's considered opinion that:

- *The conduct and results of the 1990 Trial State Assessment (TSA) provided evidence that state NAEP is technically feasible. Furthermore, the trial has generated information immediately useful to the federal and participating state governments as these bodies assess education policy.*

- *At present, NAEP excludes school-age children who are not enrolled in school. Based on the Panel's understanding that the intent of NAEP is to assess all school-age children, in 1994 NAEP should begin including a national sample of out-of-school 17-year-olds.*

- *NAEP can and should play a role at both the national and state levels because it will provide valuable and meaningful trend data at these levels.*

- *Because NAEP is the nation's only independent global monitor of educational achievement, Congress should act to ensure the continuation and protection of NAEP's status as an uncorrupted monitor of the educational achievement of the nation's children.*

Should State NAEP Be Continued and If So, in What Form?

Congress wisely decided to institute state NAEP on a trial basis. Trials have at least three positive characteristics. First, they allow for an evaluation of the technical feasibility of a program such as state NAEP before a sizeable financial commitment to the program is made. Second, trials provide information for fine-tuning a program before it is made permanent. In the case of the TSA, it appears that the program *is* technically feasible, but improvements can and should be made. Finally, trials can generate data that are in-and-of-themselves useful. For example, several of the participating states have already found the data from the 1990 TSA useful, as noted in Chapter 5.

Although the 1990 trial provided a considerable amount of information concerning the feasibility of state NAEP, the 1992 trial will conclude with evaluations completed in only two subjects, mathematics and reading, at two grade levels; there will be no trial data for other subjects such as science, geography, and history. Much could be learned from a set of 1994 trials that more closely resembles a fully implemented state NAEP. Moreover, if a 1994 TSA is authorized, trend data for fourth grade mathematics and reading would become available, along with a continuation of the trends for eighth grade mathematics. The availability of this additional trend data would allow for a more complete evaluation of the uses and value of these data by the participating states.

Furthermore, with the completion of the 1992 NAEP, there will have been no 12th grade trials. Results from the trials at the 4th and 8th grade levels cannot be assumed to generalize to the 12th grade. One consideration is that the motivation of 12th graders to participate in and perform well on NAEP may differ from that of students in the lower grades. The inclusion of 12th grade trials in the 1994 state NAEP would almost certainly provide new and valuable information. For example, the results of a 12th grade assessment would be of great interest and value to the states because this age level provides information relevant to workforce preparedness.

Finally, continuation of the trial program can reveal whether the feasibility of data collection declines as the program is extended to all subject areas and three grade levels. Only an expansion of the trials to additional subjects at all three grade levels will provide the data needed to more fully evaluate the feasibility, impact, and cost of a fully implemented, permanent state NAEP.

On the basis of its findings, the Panel believes that Congress should approve the continuation of state NAEP, but before legislating a permanent state NAEP, should authorize additional trials. Because only two subjects at grade four and one subject at grade eight will have been assessed at the conclusion of the 1992 TSA, and because expansion of the trials to more subjects at all grade levels would provide the data needed to evaluate the feasibility, impact, and cost of a more fully implemented state NAEP, the Panel recommends the continuation of the trial program in 1994, rather than the full establishment of a state NAEP program. In particular, the Panel recommends trials in 1994 in the 4th, 8th, and 12th grades in mathematics, reading, and one additional subject, such as science.

Should NAEP Provide an Assessment of the Educational Achievement of All of the Nation's School-Age Children?

NAEP assessed national samples of 9-, 13-, and 17-year-old students from its inception through 1988. In the 1984 assessment, NAEP expanded the samples to include students in grades 4, 8, and 11 as well. Beginning in 1990, new trend lines were begun for grades 4, 8, and 12, but bridge samples of 9-, 13-, and 17-year olds were added to maintain trend lines for these groups. In earlier years NAEP also included out-of-school 17-year-olds, but the difficulty of sampling and data collection combined with cuts in funding eventually eliminated this group from the sample.

At present, NAEP is an assessment of what children *enrolled in school* at the 4th, 8th, and 12th grade levels know. NCES, using data from the Census Bureau's *Current Population Survey*, reports that 14 percent of 18-year olds are neither currently enrolled in school, nor have they completed high school.[1] Other evidence suggests that the number of dropouts may be even higher. Recent data indicate that over a quarter of the nation's children drop out of school between grades 7 and 12.[2] Studies show that children who drop out have lower standardized tests scores and lower grade-point averages while they are in school.[3] Moreover, because the dropout rate has changed and continues to change across time, the composition of the groups of students enrolled in school and therefore assessed by NAEP continues to change. *These continuous changes in the composition of the school-age population call into question the validity of current comparisons of the educational achievements of 12th graders across time.*

Such validity issues are of paramount importance when making comparisons of 12th graders across time, but perhaps of more importance is the paucity of information about the knowledge and skills of that proportion of the population of children of 12th-grade age who have dropped out of school and are not currently assessed—a group of young people at high risk for crime, delinquency, drug use, and welfare status. Although it is assumed that their employability is low because they lack the requisite knowledge and skills, reliable and valid data with which to compare their educational achievement against that of young persons of the same age who are enrolled in the 12th grade do not exist. The Panel believes that it is important to know whether these children have at least some of what the Secretary's Commission on Achieving Necessary Skills (SCANS) calls foundation skills—skills that are necessary for the vast majority of jobs including reading, writing, mathematics problem solving, and reasoning skills.[4]

[1] U.S. Department of Education, National Center for Education Statistics, *Dropout Rates in the United States* (Washington, D.C.: 1990), 21.

[2] American Institutes for Research, *National Dropout Statistics Field Test Evaluation: Final Report* (Palo Alto, CA: 1991), 36.

[3] R.B. Ekstrom, M.E. Goetz, J.M. Pollack, and D.A. Rock, "Who Drops Out of High School and Why? Findings from a National Study," in *School Dropouts: Patterns and Policies*, ed. G. Natriello (New York: Teachers College, Columbia University, 1987), 53.

[4] U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills, *What Work Requires of Schools: A SCANS Report for America 2000* (Washington, D.C., June 1991).

Providing a reliable picture of what all school-aged children know has been the stated aim of NAEP since its inception. Valid comparisons of assessment results across time require that samples of 17-year-olds who are no longer enrolled in school be assessed routinely. The Panel recommends that beginning with the 1994 assessment, a national sample of out-of-school 17-year-olds be assessed. This approach would be less expensive than testing state samples, but should be supplemented by giving each state participating in the TSA the option of paying for the assessment of a similar sample of its out-of-school 17-year-olds.

In earlier years, NAEP assessed young adults in addition to assessing 9-, 13- and 17-year-old school children. These informative assessments of young adults were discontinued due to declining NAEP budgets. A National Assessment of Young Adults (26-year-olds) would nicely complement what the nation will learn from NCES's ongoing National Adult Literacy Study (NALS). The NALS focusses on far more practical skills (e.g., reading a bus schedule) than NAEP. A Young Adult NAEP would provide information on what young adults know in areas of reading, writing, mathematics, geography, and history using the same frameworks and scales used on national and state NAEP. This in turn would provide important information on the longer term effects of our nation's schools. Also, like the proposed study of out-of-school children, a study of young adults would inform the nation about how well this group performs on the SCANS foundation skills.

The Panel considered issues surrounding the implementation of a Young Adult NAEP. The cost of such an assessment would be significant. Moreover, it would add a significant burden to the staff of NCES and its contractors, and rapid growth in a short time period could be deleterious to the operation of national and state NAEP. It is the Panel's opinion that Young Adult NAEP should receive further investigation prior to a recommendation for its implementation. Therefore the Panel will continue to examine the advisability of a Young Adult NAEP in its evaluation of state NAEP over the next two years.

Should the Prohibition Against Reporting NAEP Scores Below the State Level Be Lifted ?

Current law bars reporting NAEP results below the state level. Ranking, comparing, or evaluating individual students, schools, or school districts is explicitly prohibited. Earlier this year the National Assessment Governing Board (NAGB) recommended to Congress that the current prohibition against reporting NAEP data below the state level be lifted to allow reporting at the district and school levels. (NAGB at its November 1991 meeting, however, recommended against the use of NAEP at the individual level.) The Council of Chief State School Officers (CCSSO) has recommended that the prohibition against reporting data at the district (but not the school or individual) level be lifted. *America 2000*, the Administration's education strategy for the rest of this decade, calls for the use of an individual NAEP-like test as an interim national examination.

Because arguments exist, pro and con, with respect to whether NAEP data should be reported at the district, school, or individual levels, the Panel commissioned papers for[5] and against[6] lifting the prohibition. The Panel has carefully examined each of the positions in making its recommendation about the status of the prohibition.

Arguments in Favor of Lifting the Prohibition

If the prohibition were lifted, states, districts, and schools would all have access to an instrument capable of producing higher quality data on the educational achievement of their students than is currently available to them from any commercial test publisher.

The NAEP subject matter frameworks are based on efforts to represent state-of-the-art content in the areas of reading, writing, and science. As a result, NAEP offers the potential district and school site users methods of assessing student performance in core subjects that reflect contemporary, research-based views of these subjects. Furthermore, NAEP's background data could provide important diagnostic information, allowing a school or district to describe its teacher work force, its instructional practices, or the distribution of instructional resources available to various groups of students. This information would undoubtedly be of tremendous value to local program analysts and managers.[7]

Because of our country's historically decentralized educational system, data collection has evolved as a combination of national, state, and local programs. Those who favor the use of NAEP at the school and district level argue that if NAEP were available for use at these levels it would allow for the vertical integration of educational achievement data within states.[8] That is, comparable achievement data would be available at the national, state, district, and local levels.

These are important reasons for recommending that NAEP be available for use at the school and district levels, but the Panel believes there are even more compelling reasons for not lifting the prohibition.

Arguments Against Lifting the Prohibition

First, if the prohibition is lifted, it could discourage important state and local initiatives in assessment. A number of states and districts have developed innovative, high quality testing programs. Furthermore, these assessments are tailored to meet the

[5] Ramsay Selden, "The Case for District- and School-Level Results from NAEP," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

[6] Walter Haney and George F. Madaus, "Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting Below the State Level," in *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* (Stanford, CA: National Academy of Education), 1992.

[7] Selden, "The Case for District- and School-Level Results from NAEP."

[8] Ibid.

states' or districts' specific needs. State and local assessment efforts have led the way in showing that it is possible to set ambitious curricular goals and to construct cutting-edge assessment techniques to measure them. Arizona, California, Connecticut, Kentucky, Maryland, and Vermont are all excellent examples of such efforts. It is quite plausible that if NAEP were made available at the local level, many state, district, and local testing programs might be eliminated in order to save money. If these sources of innovation were lost, not only would education as a whole suffer, NAEP itself would suffer substantially because much of its innovation has derived from the work of exceptional state assessment programs. It is also important to encourage the innovations of *local* policymakers, teachers, and school administrators: Lifting the prohibition would almost certainly discourage such efforts.

Second, there is a growing body of evidence that the higher the stakes associated with a test, the more likely it is that teachers will teach specifically to it.[9] Furthermore, research suggests that teaching to the test may eventually erode the usefulness of the test.[10] Performance on a regularly scheduled high-stakes examination may not generalize to performance on other tests covering the same content areas for which the students have not been specifically prepared. In other words, by preparing the students for the specific content and format of the high-stakes test, teachers may boost performance on that particular test without enhancing achievement beyond its scope.

Finally, if NAEP were used at the local level and became a high-stakes test, item security would almost certainly become a problem. NAEP uses a sampling design that depends heavily on safeguarding the items for reuse. In order to use the current design for local level NAEP, many more items would have to be replaced at each assessment cycle. This would not only be very time-consuming and expensive, it could also impact NAEP's ability to produce valid trend lines.

NAEP has been and should continue to be a *global* monitor of educational achievement. Its historic role as an independent auditor that neither is unduly influenced by particular instructional or testing practices nor shapes day-to-day classroom activities would be lost if TSA results were to be reported at the district, school, or student levels. In addition to compromising NAEP's value as an indicator, lifting the prohibition could detract from local innovations that attempt to integrate assessments with curricula. Therefore, the Panel concludes that the prohibition against reporting NAEP data below the state level should be continued.[11]

[9] Haney and Madaus, "Cautions on Removing the Prohibition Against the Use of NAEP Test and Data Reporting Below the State Level."

[10] D. Koretz, R. L. Linn, S. B. Dunbar, and L.A. Shepard, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests," Annual meeting of the American Educational Research Association and the National Council on Measurement in Education, in D. M. Koretz (ed.) (Chicago: Rand Corporation), 1-30.

[11] Panel member Ambach dissents from this position; he is on record elsewhere as recommending lifting the prohibition at the school district level where the size of the enrollment enables the use of the same sampling as used at the state level.

What Role Should NAEP Play in Education Reform?

NAEP has a role to play in education reform at both the national and state levels. While the trend results from NAEP cannot be used to draw inferences about the causal agents responsible for changes in educational attainment, the results can nevertheless be of great use to the states in evaluating the nature of student achievement and suggesting areas for instructional attention. The trend data that NAEP collects at regularly spaced intervals can be useful to states and to the nation in helping to measure the progress (or lack of progress) toward their educational goals.

NAEP Can Be a Monitor of the Status of Education at the National Level

Both the Administration, as part of its *America 2000* strategy, and the Congressionally appointed National Council on Education, Standards and Testing (NCEST) have called for a new national examination system. The National Education Goals Panel (NEGP) has also called for a system of performance-based examinations and portfolios. Although considerable debate surrounds the precise nature of any national examination system, present proposals all agree that the examinations would be voluntary. Many view the main purpose of such examinations as providing feedback to teachers and schools on student performance on authentic tasks related to NEGP's Goal Three, which states that students should leave the 4th, 8th, and 12th grades with demonstrated competence in English, mathematics, science, history, and geography.

Because any such examinations would be voluntary, the results would *not* yield representative national results; i.e., results generalized from the participating schools and districts would be biased to some degree. The degree of the bias in the estimates would be a function of the factors determining which schools and school districts participate. Unfortunately, there is no way to estimate the degree of bias without having a separate, independent assessment based on a probability sample against which to compare. Both NEGP and NCEST have suggested that NAEP should be the independent monitor of educational achievement in the country. *The Panel strongly endorses this position*. The Panel believes the proposed role for NAEP is not only appropriate, but exemplifies the type of role NAEP can and should play in reflecting educational progress in the nation.

NAEP Can Monitor the Results of State-Level Education Reforms

State NAEP will become an increasingly powerful tool for monitoring educational progress within states. For the states that participate in both the 1990 and 1992 trials, trend data for those two trials will be available in 1992. If the trials continue in 1994 as recommended by the Panel, two-point trends for fourth grade mathematics and reading will also be available, in addition to data on eighth grade mathematics from three points in time. As some states attempt major improvements in their schools, the trend data that state NAEP promises to offer will be of considerable value in assessing their educational attainment.

Of special importance will be the information that NAEP can provide to state programs for at-risk students. The availability of trend data for and about various groups of students could provide indicators of both the need for such programs and possible contributions to student performance. Similarly, NAEP could be used to evaluate whether different segments of the student population are performing in different ways. For example, a state may want to track the average performance of students in the top and bottom 10 percent of the NAEP distribution across time to see if the gap between the two deciles is narrowing or widening. Or, where warranted by demographic and other similarities, a state may want to compare the performance of a particular group of students across time with those from another state.

Although NAEP can be used to measure changes in academic performance at the national and state levels, it is important to emphasize that *NAEP cannot be used by itself to infer that any observed changes were caused by the reforms.*[12] For example, even in cases where NAEP suggests improvement over time, such results cannot be taken as definitive evidence that the change is due to any specific reform. The changes might just as well be due to changes in the demographic composition of the state (e.g., in- or out-migration), or due to a combination of both causal and non-causal factors. The Panel's emphasis on the inability to draw causal inferences from NAEP trend data alone is driven by more than an academic perspective; it considers policy implications as well. Just as NAEP scores across time could show performance gains that are not actually the result of education reforms, it is also entirely possible that positive educational improvements that *are caused* by a statewide, systemic education reform program could be missed by state NAEP trend data. For example, if a state has a recent, large influx of foreign-born immigrants, the *positive* effects of the reform program could be offset by the initial poor performance of recent immigrants who have not been in the country long enough to acquire English proficiency nor in the state long enough to profit from its educational investments.[13]

It is also important to point out that NAEP can suggest areas in which continuing attention to instructional improvement seems most needed. As a consequence, the Panel believes that most if not all of the states participating in state NAEP will find the trend data of great utility in monitoring the changes in their education systems.

State NAEP Can Be Useful in Helping to Identify Strengths and Weaknesses in Education Systems

NAEP can be of use in helping to identify strengths and weaknesses in states' education systems as they consider reform activities. NAEP teacher questionnaires yield data on teacher characteristics and instructional practices. The 1990 mathematics assessment showed significant relationships between students' performance in mathematics and (a) the mathematics content that their teachers studied in college and

[12] Edward Haertel, "Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot be Made," *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*, (Stanford, CA: National Academy of Education), 1992.

[13] This argument assumes that states with a large population of immigrants will have a substantial number of students whose English-speaking skills are not so poor as to eliminate them from a NAEP assessment as Limited English speakers (LEPs), but whose English-speaking skills are nevertheless weak enough to result in diminished performance on NAEP.

(b) their teachers' involvements in professional mathematics activities. These are but two examples of how NAEP background data can be used to provide important diagnostic information. NAEP data allow the nation and states to describe strengths and weaknesses of their teacher workforces, their instructional practices, and the resources available to their students, and also aid in determining whether there are significant relationships between these factors and student performance. In addition, the fact that all of the states have collected data with the same instruments allow states to compare their findings with findings of other states, assuming the states are similar enough to warrant such comparisons.

NAEP results have already proven very useful for tracking the achievement of minority groups across time at the national level.[14] With the implementation of state NAEP, states can begin to construct similar trend data. The educational progress of minorities can be related to data on coursework, computer time, homework time, and similar factors. The presence of significant relationships can point to areas where further research is needed to assess which of the observed relationships are causal.

Summary

The Panel believes that NAEP results will be useful in the context of current ambitions for education reform, whether or not a proposed system of national examinations is implemented. *The Panel further believes that NAEP should be preserved as the nation's key independent indicator of how students are performing academically. NAEP has played and should play the same role for education that economic indicators have played in signaling changes in the status of the economy.* Without question, economic indicators are imperfect and can be improved. So, too, can NAEP be improved. In the meantime, just as it is important for the nation to preserve independent indicators for determining the status of its economy, it is important to preserve NAEP as the key monitor for assessing the health of our nation's education system. For NAEP to continue this role, it must not become corrupted by use at the individual, local, or district levels. NAEP designers must move cautiously in developing performance-based instruments and strategies which reflect new advances to which schools are yet adjusting. NAEP must reflect a balance between innovation and proven pedagogy. Its frameworks must lead curriculum and instruction at the same time that they reflect the outcomes of current practice. If NAEP can maintain these balances, it will play an important role as a key indicator of the status of education and the needs for changes in policy and practice at the state and national levels.

[14] U.S. Department of Education, National Center for Education Statistics, *America's Challenge: Accelerating Academic Achievement* (Washington, D.C.: 1990), 38-47.

Appendix

Synopses of Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment

This appendix contains summaries of papers that were commissioned by the Panel and served as the basis for the Panel's findings and recommendations. The complete papers are being published in a separate volume, *Studies for the 1990 Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. The Panel also commissioned a study of the validity of the 1992 reading framework and items. Because it is more relevant to the next evaluation, it will be included as part of the Panel's report on the 1992 Trial State Assessment.

A Critique of Sampling in the 1990 Trial State Assessment

Bruce D. Spencer
National Opinion Research Center
Northwestern University

The 1990 Trial State Assessment (TSA) was restricted to the eighth grade public school population in 37 states, Washington, D.C., and two territories. The subject of this investigation is the design and implementation of the sample, including issues of nonparticipation of students who were intended to participate in the TSA.[1] Specifically examined were the specification of target of estimation (if 100 percent sampling were used); the adequacy of the sampling frame, at both the school and student levels; implementation of the sample design, the uses of stratification, and variability of designs among states; poststratification; issues in merging national and state samples, particularly weighting; the efficiency of sampling design and weighting; the nonresponse, at state, school, student, and item levels; the adequacy of methods for estimating standard errors and other measures of uncertainty; the adequacy of methods for adjusting for missing data, both weighting and imputation; and the accuracy of certain statistics.

Sampling of Schools

The sampling design called for sampling about 100 public schools per state; if a state had fewer than 100 schools, then all were selected. To ensure that students within each state had an approximately equal chance of being selected in the sample, schools were selected with probabilities proportional to their estimated eighth grade enrollment. Special procedures were used for small schools. The lists of schools were based on the National Center for Education Statistics (NCES) Common Core of Data for 1988, with state samples selected by systematic stratified sampling based on urbanicity, percentage of students who were black or Hispanic, median income, and eighth grade enrollment size. This is standard stratification practice and appropriate for broad-based surveys.

In all, the sample design was the standard kind used in high quality, major national education surveys. It seems to have been executed with care and due attention to details. Different sampling designs (i.e., different stratification and different methods of handling small schools) were used in different states, and it is reasonable to expect some efficiency gains if the stratification could be based on such prior school-level test scores as statewide exam scores. Future assessments should give attention to stratifying schools within a state by test scores on statewide exams, and this stratification approach should be empirically evaluated in at least a few states.

[1] For information on students excluded from the assessment by design, see *Eligibility/Exclusion Issues in the 1990 Trial State Assessment*, also by Spencer and included in this appendix.

Participation rates were reasonable for a survey of this nature. The overall participation rate for schools was about 94 percent, although it was less than 80 percent in two states, Illinois (78 percent) and Oklahoma (76 percent). The state average participation rates of the selected students was between 92 percent and 96 percent, except in two states, where the rates were 88 and 80 percent. Substitution increased the overall participation rate to 97 percent, although the rates were as low as 86 percent in New York. Weighting adjustments for school nonresponse were implemented. The investigator conducted a simple analysis to consider the effect of school nonparticipation in the TSA in Illinois; future assessments should include evaluations of nonparticipation (nonresponse bias) at the school and student level.

Sampling of Students

A systematic sample of students was drawn from a list provided by each school selected for the TSA. A supplementary sampling procedure was used to account for students who were originally selected but transferred before the assessment occurred; however, it seems to have accounted for only 60 percent of the students it was supposed to cover. The contractor reported that 4.9 percent of the students withdrew from the sample, but the supplemental sampling procedure added only 2.9 percent, leaving 2 percent who seem to have been missed by the survey. Future assessments should give more attention to capturing transfer students.

Participation rates of selected students generally were good, ranging from 92 percent to 96 percent except in two states, where the rates were 88 percent and 80 percent. In most states, absentees comprised the major portion of the nonparticipating students in the participating schools. Make-up sessions were offered but captured only about one-fifth of the students who did not participate in the initial testing. Because participating and nonparticipating groups tend to differ on many kinds of characteristics, an analysis of the performance of absentee students on make-up tests should be conducted in future assessments. Differences in the demographic composition of participant and nonparticipant groups are also commonly observed in national educational surveys, and although weighting adjustments for student nonresponse were implemented, the magnitude of the effect of student nonresponse warrants further study in the future.

Weighting and Sampling Errors

Weighting classes were used to adjust for school-level and student-level nonresponse. The basic assumption behind the use of a weighting class is that within a class persons decide to participate or not at random; thus, within any weighting class or cell, the mean score for participants is expected to be the same as mean score for nonparticipants had they been assessed. The effectiveness of nonresponse adjustments essentially depends on how well this assumption describes reality. For the TSA, the weighting classes were based on stratification variables at the school level and demographic variables at the student level. The school-level weighting classes were appropriate; student-level weighting classes were based on characteristics that

were only somewhat correlated with test scores, and the single student-level characteristic was an age dichotomy (younger than 13 years old or at least 13). Characteristics of nonparticipating students were not recorded, so other data were not available. Given well-documented differences in scores for different gender/race/ethnic groups, as well as inevitable differences in participation for those groups, the nonresponse adjustments cannot be expected to be very effective for removing bias due to nonparticipation by students. Future assessments should include attention to collecting data for nonparticipating students for the purposes of nonresponse adjustments.

The magnitude of the sampling error (ignoring item-sampling errors and parameter-estimation errors) for state-level estimates and rankings of states was reasonable. The investigator calculated the magnitudes of sampling errors for proportions of students whose scores exceeded various proficiency levels as the coefficient of variation, a ratio expressing the sampling error in relative terms. The coefficients of variation were quite high for the proportion at the advanced level, and moderately high for the proportion at the proficient level. Acceptability of these levels depends on the needs of a particular analysis in which the estimated proportions will be used.

Eligibility/Exclusion Issues in the 1990 Trial State Assessment

Bruce D. Spencer
National Opinion Research Center
Northwestern University

The 1990 Trial State Assessment (TSA) was restricted to the eighth grade public school population in 37 states, Washington, D.C., and two territories. The survey excluded certain large groups of students, some by design (e.g., private school students) and some by chance (e.g., absentees). The focus of this investigation is on exclusions by design. Spencer considered the possible effects of variable exclusion criteria across or within states; differences between the design for exclusion and the actual field practice for exclusion; on-going data collection features that might be embedded in state assessments to permit statistical evaluation of the impact of exclusions and nonparticipation; and past research on the impact of exclusions on TSA statistics.

Who Was Supposed to be Assessed?

The design of the TSA allowed for the exclusions of students enrolled in private schools, students with limited English proficiency (LEP), and special education students with individualized education plans (IEP). Nationwide *one in six* students belongs to one or both of these groups. Because states vary widely in the proportions of students belonging to these groups, these differences could have important effects on state-by-state and state-to-national NAEP comparisons.

Private School Students

Private school students typically made up the largest excluded group in each state. Nationally, about 12 percent of eighth grade students are enrolled in private schools, but the percentage varies widely across states. In seven states, fewer than 5 percent of the eighth graders are in private schools, whereas in seven others, more than 18 percent are in private schools. In Wyoming and Utah, only about 2 percent of the students are in private schools; in Hawaii and the District of Columbia, they account for about 20 percent. The investigator conducted illustrative sensitivity analyses of the states' scores in the 1990 TSA to suggest how large the effects of excluding private school students might be. The first set of analyses was based on state mean scores; the second was based on the proportion of students in each state scoring at or above the three thresholds corresponding to basic, proficient, or advanced performance. The proportion of eighth grade students in each state who were enrolled in private schools was estimated from data compiled for the development of the NELS:88 sample by the National Opinion Research Center.

The analyses based on mean scores utilized three models, each making a different assumption about the effects on state ranks of excluding private school students. For example, the first assumption was "had private school students been tested, on average their scores would have been 0.6 standard units higher than their public school counterparts in the state." This analysis suggested that the rankings for several of the 38 states (including the District of Columbia) could change up or down by three or more ranks, and the assumption likely understates the real variation in overall state mean scores. In each of the three simulations, the effect of excluding private school students was at least twice as large as the effect of sampling error in the TSA.

Sensitivity analyses based on proportions of students meeting basic, proficient, or advanced performance standards were also carried out with three models. National estimates of the difference between the logits of the proportions of public school students and private school students meeting the various performance standards were used. These differences were assumed to describe state-level differences. For example, the first sensitivity analysis assumed that the same national public-private logit differences were present in each state as well. In all, the magnitude of the effect of excluding private school students ranged from about the same as sampling error to twice as large. The report concludes that state NAEP should include private school students as well as public school students.

LEP and IEP Exclusions

The exclusion criteria for LEP and IEP students were part of the sampling design and were implemented locally. Local administrators were told to exclude students who could not be expected to meaningfully participate but to include doubtful cases in the assessment. Attempts to compare the proportions of students classified as IEP and LEP in the TSA with other available data were hindered by the weaknesses of the available comparison data. Differences in the percentages of LEP students tended to be about 1 percent lower for the TSA than for the comparison statistics, except in New Mexico, where comparison data put the percentage of LEP students at 16.4 percent and the TSA put it somewhat under 2 percent. The definition of LEP may have been applied inconsistently within states as well as across states, because the operationalization of the concept of "LEP" was not fully specified in the design. The discrepancies for IEP ranged from about 3.5 percent lower to 3.5 percent higher in the TSA, averaging about 1 percent lower in the TSA. As with LEP, operationalization of the concept of "IEP" was not fully specified in the design.

Spencer compared the TSA and national NAEP on the proportion of test-takers who were classified as IEP or LEP and found that the TSA included roughly twice as many test-takers classified as IEP or LEP as national NAEP did. The discrepancy implies that the TSA and national NAEP either (a) employed different criteria for classifying the students as IEP/LEP or (b) excluded students who were "really" IEP/LEP at rather different rates. Explanation (b), if valid, would be cause for concern because it implies that the TSA was more stringent in testing IEP/LEP students than national NAEP has been: either the TSA is too stringent or national NAEP is too lax.

Comparisons of the monitored and unmonitored sessions regarding the proportions of tested students who were classified as IEP or LEP were also made. Differences

generally were minor, although in some states the IEP differences reached 2 percent. The LEP differences were smaller and not consistently higher or lower.

The average score for IEP/LEP students who took the test was lower than the average for the other tested students. Several sensitivity analyses were performed to provide order-of-magnitude estimates for how much lower scores might have been if all IEP/LEP students in selected schools had participated. Results indicated that if it had been practical to include all of the IEP/LEP students in the assessment, the states' scores and rankings would have changed moderately.

Combined Effect of the Exclusions

An additional sensitivity analysis was run to estimate the differences in the mean standard scores and ranks of states that might have occurred if both private school students and the excluded IEP/LEP students had been tested. The average change in state mean standard scores in the simulation was observed to be 0.12, and the average change in ranks was 1.9, with some states' rankings changing by as much as seven or eight. In all, the effects are much larger than the magnitude of sampling error.

Who Should Be Assessed?

The current strategy seems based on an overly simplistic "cost-benefit analysis," which excludes IEP/LEP students because it would be relatively expensive to assess them. NAEP data are currently interpreted by almost all users with the unstated (and faulty) assumption that the excluded IEP/LEP students have the same mean scores as the assessed students. Consideration should be given to including all IEP/LEP students in the tested population. IEP/LEP students can be effectively represented in the assessment through a variety of cost-effective methods. For example, the IEP/LEP students who are difficult to assess could be sampled at much lower rates than other students, involving less data collection and more statistical modeling than is currently done, but resulting in increased accuracy. However, if exclusions are to continue, more study should be given to the formulation of consistent and valid criteria for exclusion and to the implementation of the criteria.

Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment

Donald H. McLaughlin, Frances B. Stancavage, Jay G. Chambers,
Elizabeth Hartka, and Kadriye Ercikan
American Institutes for Research

The purpose of this investigation was to assess (1) the feasibility of the full implementation of the Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP), in terms of state and local costs, and (2) the impact of state and local variations in implementation of the assessment procedures on the validity of the results. The aims were to provide both an independent assessment of potential threats to the credibility of the state-level data, as well as to identify specific areas for the refinement of the TSA in the future. The 1990 TSA was carefully planned and implemented, with substantial training of local administrators. Nevertheless, the investigators found that refinements in several areas should be considered for the TSA implementation in 1992.

Eight data sources were used for this investigation: (1) performance data on the TSA sample of 100,843 assessed students, plus a matched NAEP comparison sample of 2,487 students generated by the NAEP contractor; (2) quality control monitor data; (3) excluded student data on 5,735 individualized education plan (IEP) and limited English proficient (LEP) students who did not take the test (many IEP and LEP students did take the test); (4) school questionnaire data and school substitution information; (5) state coordinator survey data; (6) a sample of administration schedules provided by the NAEP contractor and subcontractors; (7) an American Institutes for Research (AIR) survey of state testing directors; and (8) notes from AIR observations of TSA training and administration sessions. To assess validity, both frequencies of irregularities (i.e., any variations that were recorded) and comparisons of performance related to these irregularities were computed. To make performance comparisons, the investigators felt that it was equally important to identify correlates of increases in variance as it was to identify correlates of mean differences.

Overall Validity Estimation

First, a comparison was made between monitored and unmonitored sessions, and only very small differences were found in either mean levels or variances, using either raw or adjusted scores. This is important because most of what is known about variations in session administration and their impact on performance is based on monitored sessions only. Second, a comparison was made with the NAEP comparison sample generated by the contractor to match the overall TSA sample. Because the comparison sample overrepresented minorities, it was essential to use adjusted scores in the comparison. The aggregate average TSA performance was .12 standard deviations higher than the aggregate average comparison sample performance, on a measure with a nominal population standard deviation of 1.00. Thus, the aggregate average based on the matched comparison sample was higher than the averages of 10 states and lower than the averages of 30 states in the TSA sample.

Participation Rates

Approximately 6 percent of all originally invited schools refused to participate in the TSA, although in two states the proportion approached 20 percent. Schools that refused tended to be smaller and to have fewer black students than other schools. Substitute schools were successfully matched to refusing schools on demographic factors. Performance scores in substitute schools were only .02 standard deviations higher than in cooperating original schools (.10 higher without the race and sex adjustment, however). This level of difference would not be expected to account for any significant change in state ranks.

Initial examinations of data on the 5,735 students explicitly excluded because of severe disabilities or English language deficiencies revealed no substantial factors that might account for performance variations. Numbers of these IEP and LEP students excluded in individual states ranged from 48 (4 percent) to 244 (9 percent) across the 37 states, the District of Columbia, and 2 territories. Of 3,915 sessions, only 119 had more than 5 excluded students and of these, only 22 had more than 10.

In addition to explicitly excluded students, there were students who were no longer at the school (withdrawn) or who were absent (and did not have a makeup session). It is impossible to know how these students would have scored, but evidence was found relating numbers of withdrawn or absent students to scores of those present for the assessment. A significant *negative* correlation, -0.27, was found between the number of students withdrawn from a session and the average performance score of those present in that session. Virtually the same negative correlation was observed between numbers of absent students and average scores. The most straightforward interpretation of these results is that schools with high rates of student absences and departures are also schools with lower mathematics-achieving students. One state differed from all others in the frequency of absences, and the presentation of results for that state should be qualified by the large absence rate.

Preassessment Variations

Quality control monitors (QCMs) made preassessment checkup calls to all local session administrators a few days prior to the assessment. In the majority of cases, everything was being carried out in accordance with plans at the time of the call, and performance scores were not significantly related to deviations from these procedures. Nevertheless, these calls were important because the responses correlate, in monitored sessions, with QCM ratings of the administrator's performance. Also, in many cases the teacher questionnaires had not been distributed, and the return rate for this rich source of contextual information would undoubtedly have been lower without the reminder.

Session Administration

Irregularities in session administration affect validity primarily by creating especially favorable, or especially unfavorable, testing conditions for particular students. For most potential problems, information on the frequency of occurrence is available on the forms filled out by the QCMs. However, because they were instructed to intervene to correct observed problems wherever possible, there is only limited opportunity to assess the probable magnitude of the impact on student scores in the absence of such intervention. Among the factors under the control of the local administrator, the most frequent irregularities were errors in reading the script that contained the test instructions, and errors in distributing and collecting the calculators between test sections. Errors in three or more script sections (out of five) were observed in 13 percent of the monitored assessment sessions, and errors in calculator distribution were observed in 11 percent of the sessions. Sessions with errors in three or more script sections had significantly lower student performance. No decrement in student performance was found to be associated with errors in calculator distribution, but these results might have been different if the QCMs had not intervened in virtually all instances of distribution error. After errors in script reading, the single factor most clearly associated with lower student performance was student disruption. Fortunately for the integrity of the state-by-state comparisons, these and other session administration irregularities tended to be distributed randomly over the participating states, with no state as a significant outlier.

Incentives

In 20 to 25 percent of the monitored sessions, administrators met with students prior to the assessment, in 15 percent of the cases they gave them a pep talk, in 9 percent they gave students an incentive to participate, and in 5 percent they gave special instructions in using calculators. Generally, students who were given incentives, a pep talk, or special instructions performed more *poorly* than others—a reminder of the difficulty of making causal inferences from data such as these. Examination of the effects of incentives is continuing.

Sessions were conducted by counselors, teachers, principals, and other educators. Scores in sessions conducted by principals, vice principals, and superintendents were found to be higher than in other sessions. The most straightforward interpretation of this finding is that apparent position of power of the test administrator may have affected students' perceptions of the importance of good performance on the test.

State and Local Costs for the TSA

Rough estimates were gathered of the amount of effort involved in the various tasks required for participation in the TSA. Most state testing directors perceived their costs

generally to be in line with their expectations. Retrospective cost estimates varied widely among states because of the many ways in which components of the cost could be inadvertently overlooked (e.g., the time spent by mathematics teachers to fill out questionnaires). An initial analysis of costs by function suggested that the total costs for a state's participation were probably between $90,000 and $210,000, most of which were incurred at the local level. Most state testing directors felt that the assessment was worth the cost, and the number of participating states is expected to increase in 1992.

Overall, the initial evaluation of the TSA indicates that the state NAEP is probably feasible and probably will provide the basis for reporting valid state-by-state results. A number of issues and concerns were raised, however, by the patterns observed in the data, and these should be addressed. First, monitoring of sessions provided valuable data and probably avoided many problems, and it should be continued. Second, issues such as the differences between the TSA scores and the national NAEP scores need further investigation, especially as they relate to student motivation during the assessment. Finally, training of local test administrators should emphasize, even more, the value of uniformity of administration procedures. Although the administration was of high quality, the infrequent variations that did occur were often associated with lower average performance scores.

The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis

Edward A. Silver, Patricia Ann Kenney, and Leslie Salmon-Cox
Learning Research and Development Center
University of Pittsburgh

The investigators examined the consensual process that produced the mathematics objectives framework used to select and develop items for the 1990 NAEP Trial State Assessment (TSA) and analyzed the validity of the framework and the items themselves. Data sources for the examination of the consensual process included archival records and other written materials about the development of the framework and, more importantly, semi-structured interviews with key participants: 10 state representatives who participated in the consensual process and 18 others who were either members of the Council of Chief State School Officers (CCSSO) committee or consultants to the consensual process. An expert panel of eight mathematics education specialists representing different constituencies and perspectives was assembled to scrutinize the 1990 framework and items.

The 1990 NAEP Mathematics Objectives Framework

Almost everyone interviewed or consulted agreed on the adequacy of the 1990 NAEP mathematics objectives framework. It apparently captured the spirit of change in mathematics education at the time it was written, yet preserved enough traditional elements to represent a reasonable compromise between the visions of educational reformers and the realities of educational practice. The investigators also found virtual unanimity in the view that the framework must be considered as evolutionary, not fixed for long periods of time. The framework must be allowed to evolve to reflect changing conceptions of mathematics education in the nation, especially during these times of field-based reform. As the framework is modified, a consensual process needs to be maintained so that the framework can remain consistent with the views of education leaders at the state and national levels. Opinions vary on the proper length of the life cycle of a given framework. It may be impractical and unwise to alter it with each assessment, but a regular expectation of changes to the framework and a plan for making those changes are both needed. Adequate time should be built into the revision cycle so that consensus can be maintained.

Terms used to frame the assessment and report the results need to be clearly defined. It became apparent that even the eight members of the expert panel who had been carefully informed about NAEP's definitions of the terms *conceptual understanding* and *procedural knowledge* had a great deal of difficulty reaching agreement about the way those definitions should operate in classifying items. Even greater confusion can be expected when these terms are used by broader segments of the educational

community and the general public. The frequent disagreements could have been the result of miscommunications and misunderstandings, or they could have resulted from fundamental differences of opinion about the mathematical proficiencies being tested. The former explanation certainly points to a need for increased clarity in defining terms. The latter case indicates a deeper, more substantive problem that goes beyond the need for clear definitions of terms, and that needs to be addressed both by those who report and interpret the results of the assessment and by those who design future assessments.

The expert panel and some of those interviewed by the investigators expressed some concern that using a matrix to guide test-item development fosters the notion that mathematics should be compartmentalized into discrete topics, instead of being thought of as an integrated whole. The matrix conceptualization conflicts with contemporary views of mathematics as a rich set of interconnected ideas and can lead to confusions about item classification.

The 1990 NAEP Mathematics Items

NAEP is designed to provide comprehensive coverage of topics at each grade level. Over the history of the assessment, the number of mathematics items has decreased substantially, due to a combination of economic factors and modifications in test design. The investigators found reason to be concerned about the adequacy of the number of items on the test to provide an adequate picture of mathematical performance on the broad array of topics at a given grade level. In many cases, inferences will be made about the understanding or competence of the nation's students on some mathematics topics on the basis of performance on a very small number of items. Either the number of items needs to be increased or the test needs to be altered to be less comprehensive. Information currently available about the 1992 NAEP mathematics assessment suggests that the number of items has been increased for the next assessment. Simply adding more items will not automatically improve the test: attention must also be given to issues of test design and item format.

The expert panel indicated that varied task formats are needed in the assessment. The format used for most of the 1990 assessment items was multiple choice, but the expert panel rated the few open-response items as being much more representative of current thinking about important objectives of mathematics education. Few innovative items were included in the state-by-state test. A broader range of tasks that measure a wider variety of mathematical proficiencies and performances should be developed and administered. However, increasing the number of open-response tasks will also increase the importance of attention to valid scoring of students' responses. The extent to which open-response mathematics tasks measure students' knowledge and ability with respect to language and communication rather than mathematics will become a more salient issue.

Issues Beyond the 1990 NAEP Mathematics Assessment

The investigators considered three issues that go beyond the 1990 NAEP mathematics assessment: (1) matching development and reporting, (2) focus and depth, and (3) time and planning. Each is summarized below.

Matching Development and Reporting

The improved alignment of the approaches used for test development and those used in reporting results is a major issue for future development of the NAEP mathematics assessment. The framework identifies content and ability categories and is used to create a matrix; the matrix is used to write the test items. However, the framework appears to have little importance in framing the reporting of the results. Current reporting is based on proficiency scales which are derived post hoc on the basis of student performance on the test. The proficiency scales were not used to guide the item-development process and should be examined more carefully with respect to the 1992 NAEP mathematics assessment. The mathematics education community is concerned about the wisdom of using scale scores in NAEP mathematics reports because such reporting contributes to the erroneous view of mathematics as unidimensional. Scale scores tend to direct more attention to issues of content than to issues of abilities or processes, and the post-hoc descriptions of scale-score levels lack a conceptual coherence that could help guide new task development or the improvement of instructional practice.

Focus and Depth

The investigators concluded that information obtained from NAEP would be more valuable if the design could be modified to allow greater focus and depth. The design requires interim independence and excludes clusters of related items. Some clusters of related items were designed for earlier versions of NAEP, but these clusters are no longer part of the assessment. The current broad sampling of the entire mathematics curriculum could be supplemented by a more intensive examination of "focus" topics. These topics would receive greater attention in the assessment and could be used to assess students' proficiencies and knowledge, allowing development of detailed accounts of what the nation's students "know and are able to do." If varied across assessments, the focus topics could provide an additional measure of changes in student achievement over time. If testing time and money did not permit this, an approach that concentrated on focus topics in each assessment could be considered, varying the topics systematically across assessments rather than trying to assess all areas each time.

Time and Planning

A recurring theme in discussing the 1990 NAEP mathematics assessment was the lack of adequate time to complete the complex tasks required. Given NAEP's importance as an information source about students' mathematical knowledge and proficiency, it is important to include knowledgeable persons from the field of mathematics at all steps of the process. Efforts were made to involve such people, but the value of their participation was limited by the lack of time for reflection and careful consideration. Planning must begin well in advance of future assessments. A group of experts in mathematics education and assessment should be convened as soon as possible and empowered to consider alternatives for NAEP design, content, task format, and scoring criteria. This group can increase the value of the information obtained from this important national endeavor and give coherent, sustained, and reflective attention to monitoring and actively shaping the NAEP mathematics assessment over the next decade.

The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns

Robert L. Linn and Lorrie Shepard
University of Colorado at Boulder
and Elizabeth Hartka
American Institutes for Research

The release of the eighth grade mathematics results from the 1990 Trial State Assessment (TSA) provided the first comparative results on a common achievement measure for representative samples of the TSA participants. The purpose of this study was to determine the degree to which the relative standing of a state is sensitive to the choice of scales, descriptive statistics, or groups of students within a state for which results are reported.

Factors Influencing Comparisons

Comparative results in program evaluations have shown that the outcome can be quite sensitive to the choice of test content and the relative weights assigned to components of composite scores. The choice of metric (e.g., percent correct, proficiency based on an item-response theory scaling) and the choice of descriptive statistics can also alter the relative standing of group averages.

Results for relevant subpopulations, such as racial or ethnic groups, or parental socioeconomic status, can also yield comparative results that seem to contradict overall averages. Simpson's paradox results from disparities from one population to the next in the proportion of people in a subpopulation. It demonstrates that it is possible for state A's average to be higher than state B's average for each subpopulation (e.g., rural-urban, black-white-Hispanic) when the overall average for state B is higher than the overall average for state A. Results from the TSA illustrate Simpson's paradox: for all students, Texas had an average proficiency two points lower than Rhode Island (though not significantly different). However, for each of the three racial-ethnic groups with sufficient sample sizes to permit reliable estimates, Texas students had higher average proficiency than their Rhode Island counterparts (by 8 points for white students, by 9 for black students, and by 17 for Hispanic students).

When the results of the 1990 NAEP mathematics assessment and TSA were released, reporters were urged, both in oral briefings and in the written materials, not to use the results to produce a simple ranking of states. Nonetheless, the simple rank order based on average overall mathematics proficiency appeared in a variety of news

accounts of the TSA results. This investigation focussed on rank order, not as an implicit endorsement of the use of a simple ranking of states, but to provide support for the need to look beyond any single rank ordering.

Method

Sensitivity analyses were conducted to investigate the degree to which the rank ordering of states varies as a function of the definition of scales, the metric used for reporting, the choice of statistics, and the analysis of subpopulation results used to characterize the performance of a state. Four sets of data were used: (1) overall composite scores and content-specific scores on the TSA, (2) items broken down into three mathematical process dimensions in the assessment framework, (3) categorizations based on a comparison of items to the National Council of Teachers of Mathematics (NCTM) curriculum standards[2], and (4) scores computed separately for multiple-choice items and for open-ended items included in the assessment. Data from all 37 states participating in the TSA were included, but not the results from the District of Columbia, Guam, and the Virgin Islands.

Five plausible values for each content area were provided by the TSA contractor; they were rescaled to the final proficiency scale metric using a linear transformation. Percent correct scores and logits based on the percent correct scores were computed. The rank order of states using these scores was compared to the rank order using the plausible value scoring technology.

Results

Using each of the four sets of data described above, the investigators compared the resulting rank orders to those based on the state averages on the overall mathematics proficiency composite. The relative standing of a state was altered relatively little for reporting metrics including the NAEP scaled composite scores, scaled content subscores, process subscores based on either the NAEP framework or the NCTM themes, or simple percent correct scores. However, several other factors that were investigated had substantial effects on the relative standing of a state.

States that have high or low standing in terms of average student achievement may rank quite differently in terms of the performance of their highest or lowest achieving students. For example, with ranking ordered on the 10th percentile rather than the mean, Maryland and New York would have their rank orders decrease by seven or more ranks; Kentucky, New Mexico, and West Virginia would move up seven or more ranks. With ranking ordered on the 90th percentile instead of the mean, Connecticut, Maryland, New Jersey, and Virginia would have their rank orders increased by six or more ranks; Arkansas, Indiana, and Oklahoma would have theirs decreased by six or more ranks. All of these changes reflect differences in the within-state variability of performance and underscore the need to consider more than central tendency when comparing states. The fact that the 90th percentile is 95 points higher than the 10th

[2] See *The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis*, by Silver, Kenney, and Salmon-Cox, in this volume.

percentile in Hawaii, while the corresponding difference is 66 points in Montana, conveys comparative information about the mathematics performance of students in those states that is not revealed by a comparison of their means. Correlations based on percentages of students within a state scoring above anchor points of 200 and 300 also reflect some sizable shifts.

Two ways of defining subpopulations were also investigated: racial-ethnic groups and samples defined by students attending the top and bottom third of the schools within each state. State averages for the total sample on the overall mathematics composite were compared to state averages based on the state samples of white students, black students, and Hispanic students where the sample sizes were judged by NAEP to be adequate to produce reliable estimates. All 37 states had sufficient samples of white and Hispanic students to produce estimates; only 28 had sufficient samples of black students for estimates. The rank-order correlations between the total sample averages and the averages for white, black, and Hispanic students are smaller than those reported for other analyses. For example, for white student averages,the correlation is .85, and for Hispanic students, only .51, reflecting considerable differences in the standing of states in terms of average performance for all students and the average performance for Hispanic students.

The investigators also found that the rank orders of the average mathematics achievement of white, black, and Hispanic students can be quite different. When the performance of students in different racial-ethnic groups used for comparisons and states are divided into quartiles, major shifts in ranks are found. For example, only two of the states that rank in the top quartile based on the performance of black students also rank in the top quartile based on the performance of white students. One state that is in the top quartile for black student performance is in the lowest quartile based on white student performance. The difference in means on the overall mathematics composite score for white and black students ranged from a low of 20 points on the NAEP scale in Kentucky to a high of 43 in Wisconsin; the percentage of black students in the two states is approximately the same (9% and 8% for Kentucky and Wisconsin, respectively). A similar range of differences in means between white and Hispanic students was observed: Florida, for example, where 17% of the students were Hispanic, had a mean difference of 19 points on the NAEP scale; Connecticut, where 10% of the students were Hispanic, and Pennsylvania, where 5% were Hispanic, had a mean difference of 42 points. Among the possible reasons for the group differences are migration patterns, concentrations in urban areas, differential poverty rates, and identifiable subpopulations within the larger groups being compared; however, the marked contrasts between states in the magnitude of the group difference in average achievement also raise questions about differential access to education.

Rank-order correlations of average overall mathematics composite scores for the total sample and for students in the top and bottom third of schools within a state show substantial correlation, but the rankings of five states would change by six or more ranks, depending on whether ranks were based on students in the top or bottom third of schools within a state. The investigators' analyses underscore the importance of going beyond simplistic comparisons based on average overall achievement.

A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report

Frances B. Stancavage
American Institutes for Research
Edward Roeber
Council of Chief State School Officers
George Bohrnstedt
American Institutes for Research

The ultimate test of the value of the Trial State Assessments is whether they produce data that are useful to educators and policymakers. This study, which is still ongoing, utilizes interviews with state, local, and national educators and policymakers, supplemented with case studies and analyses of newspaper clippings and reporting artifacts, to answer questions such as the following: How did the states handle the reporting of results? Which of the data and reports released by NCES were judged useful, and which were not? How and to what extent are the results being used? Ultimately, will TSA results cause local districts and states to change the curriculum provided to students or instruction in the area tested? Will local testing programs change in response to the TSA? And, if any or all of these changes occur, will mathematics educators and others view the changes as positive or negative?

Methodology

Periodic telephone interviews are being conducted with (1) testing directors and mathematics specialists in each of the 50 states and the District of Columbia, (2) individuals and organizations involved in the development of the 1990 mathematics content objectives and framework, (3) federal congressional staff concerned with education, and (4) an additional sample of local and state educators and policymakers drawn from eight states that participated in the 1990 TSA and two that did not. For the state and national samples, the same respondents are being interviewed at each data point wherever possible. On the other hand, because the primary focus of the local interviews is to gauge awareness of the TSA results among the individuals responsible for local educational policy and implementation, new samples of local respondents are being drawn for each data cycle.

During April-May 1991, just prior to the release of the 1990 TSA results, interviews were held with 302 respondents, including 48 state testing directors, 48 state mathematics specialists, 23 individuals connected with the objectives development process, 19 federal congressional aides, 54 additional state educators and policymakers, and 148 local district personnel. (The overall response rate was 77 percent.) At that time, baseline measures were obtained on knowledge and attitudes

about NAEP and the TSA, preferences for alternative measures of student achievement, and (for state and local respondents) the perceived adequacy of their own state mathematics programs. In addition, the state testing directors and mathematics specialists were asked to describe their states' plans for the dissemination of the 1990 results and to explain the process by which these plans were derived.

A second round of interviews was held during July-September to gather more detailed information on the ways in which the states handled the reporting of results and to assess the immediate, short-term impacts of the results. Because the interviews were held during the summer holiday period, local respondents were excluded from this data collection cycle. Two-hundred-nineteen respondents were interviewed for an overall response rate of 90 percent.

In addition, project staff (1) attended the June and September press conferences at which the 1990 TSA data were presented by NCES and NAGB, (2) reviewed press releases and related materials solicited from the state-level interviewees, and (3) analyzed the content of press clippings from the top 50 newspapers in the United States as well as smaller dailies in the ten sample states. More interviews are planned for January-February 1992 and for approximately one year later. In the latter cycle, the study will also begin to track the impact of the reporting of the 1992 TSA.

After the January-February interviews, efforts will be made to identify and document, in a case study format, more of the gestalt surrounding the release and utilization of the 1990 TSA results. Case studies will be prepared for the eight TSA-participant states selected in the state sample, plus as many as seven other states or localities in which instances of particularly good or particularly bad reporting or data use come to the attention of the project staff. To facilitate this process, additional written input will be collected from identified informants at the case study sites.

Interview Findings

State Testing Directors and Mathematics Specialists

Seventy-eight percent of the states (including all but one of the states that participated in the TSA) prepared a press release discussing the TSA results, and nearly half held a press conference. The printed reports and other materials prepared by NCES were widely used as background material, and to assist in developing state-level communications. The data elements that states found most useful were the levels of performance (1) on the total score, (2) in the separate mathematics content areas, (3) for various student subgroups, and (4) according to various instructional practices.

States reported that the release of results provoked discussions of them, particularly in the state department of education, the state board of education, and with state-level educational organizations. The overall impact of the results on education in their state was reported as positive by 45 percent of the respondents, mixed by 45 percent, and negative by only 3 percent, with 8 percent having no opinion.

By late summer, when the most recent interviews were done, very few of these state specialists reported seeing educational reforms motivated by the TSA results or based on TSA findings. However, a number indicated that the results provided strong support for reforms already underway, and others mentioned that the impact of the state assessments could be expected to increase once successive rounds of testing had occurred and the states could review their performance over time.

Other State Educators and Policymakers

The supplementary 10-state sample of state educators and policymakers (which included governors or staff, legislators or staff, state superintendents, and presidents of state organizations such as the mathematics organization, parent organization, and teachers' union) generally concurred with the sentiments that the TSA results were useful and that the overall impact of the results on education in their state (to the extent that it could be known) was positive or mixed. When interviewed in late summer, however, they were considerably more likely than the state testing directors and mathematics specialists to say that it was too early to tell what the overall impact would be. These respondents identified as most useful the data on the overall performance of students in the state and the nation, the relationship between student achievement and the instructional variables, and the state comparisons.

Federal Policymakers

Eighty-three percent of the federal congressional staff interviewed rated the 1990 TSA as at least moderately successful in providing information that will help educators and policymakers improve U.S. education, while 6 percent said it was not successful and 11 percent did not know. The focus on the overall performance of the states was most frequently cited as a strong point of the program, while a number of the respondents noted problems that they felt limited the utility of the TSA, including aspects of the test design and the sampling.

The group was split in their opinion about whether the TSA results would result in significant changes in mathematics instructional or assessment programs—about a third said "yes," another third said "no," and the remainder indicated that it was too soon to tell. Almost all (83 percent), however, felt that it was at least somewhat important to be able to measure state performance against a common standard like the TSA, and a similar number indicated that there is no other assessment program that can better monitor the quality of education than NAEP.

Participants in the Objectives Development Process

This group of respondents included individuals who had served on the committees that developed the objectives and framework for the 1990 mathematics assessment as well as current presidents of organizations that had sent representatives to those committees. They were somewhat more critical than the state-level respondents or the congressional staffers, with only 68 percent rating the 1990 mathematics

assessment as at least moderately successful in providing information that could improve U.S. education. Twenty-eight percent said that it was not successful, while 4 percent had no opinion. The new mathematics framework, the state comparisons, and the attention given to problem solving, open-ended exercises, and the use of the NCTM *Standards* were most often cited as strong points of the program; problems that limited the utility of the assessment were identified as the reporting of irrelevant comparisons, the lack of instructional value, and the narrowness of the questions used.

More than half of this respondent group stated that they did not expect the NAEP results to affect state instruction or assessment. Although they were in agreement with the congressional staffers in feeling that the ability to measure state performance against a common standard was at least moderately important, more than half felt that some other testing program could do this job better.

General Issues in Reporting of the NAEP Trial State Assessment Results

Richard M. Jaeger
University of North Carolina at Greensboro

Macro-level concerns about reporting state National Assessment of Educational Progress (NAEP) results to various audiences are the subject of this paper. A central focus is reporting that fosters valid interpretations of NAEP results and minimizes unwarranted and inaccurate interpretations. Because state-level policymakers were consulted to identify their reporting needs before the Trial State Assessment (TSA), Jaeger directed most of his attention to the reporting of TSA results to lay audiences.

NAEP Reporting in Comparison to Appropriate Practices

Jaeger reviewed public reporting in light of the appropriate practices specified in the 1985 *Standards for Educational and Psychological Testing*. Of 16 standards pertaining to score reporting, six obviously apply and were considered in light of the reports of NAEP TSA results.

Standard 2.10, which stipulates that standard errors of measurement should be reported at critical score levels, was met. For data tables and figures reporting the estimated proportions of students whose TSA achievement is at or above the four emphasized "anchor levels" of the NAEP scale, standard error estimates were included. Standard 4.1, that scales should be described clearly, was not met. Because the NAEP scale was developed through a complex use of matrix sampling and item response theory, it is difficult—perhaps impossible—to communicate it clearly to policymakers and the public. Standard 4.2, that scales should be consistent with intended purposes, was problematic because the NAEP TSA has several purposes. Some are facilitated by the unitary cross-grade scale used to report student performance; other purposes may perhaps be better served by separate, within-grade performance scales, warranting further discussion of these issues. Regarding Standard 6.8, that persons releasing test results to the media should provide information to facilitate interpretation of test results, the investigator found clear evidence in the public reports of the 1990 NAEP TSA that this criterion was not met.

The last two standards, 3.16 and 15.10, provide the major focus of the analyses. Standard 3.16 states that score report forms should facilitate appropriate interpretations; Standard 15.10 specifies that those responsible for testing programs should provide appropriate interpretations for students, parents, legal representatives, teachers, or the media. To examine the reporting, the investigator analyzed the press briefing materials prepared by the National Center for Education Statistics (NCES) and its contractor, Educational Testing Service (ETS); the Executive Summary of the NAEP composite report; and reports in three national newspapers and two local newspapers. The overall findings were that the reports intended to facilitate public communication

of NAEP TSA results did not meet their public reporting objectives and that newspaper reports were replete with misinterpretations and erroneous conclusions.

Seven Inherent Problems and Strategies for Amelioration

Jaeger based his analysis on seven convictions about the problems inherent in reporting NAEP TSA results to the public. For each conviction, he provided evidence of the problem and offered one or more recommendations for alleviating it. Each conviction, its evidence, and the recommendations are summarized below.

1. *The press will demand simplicity in reporting, regardless of the validity of the inferences likely to follow from simple reports.* This demand was evident in the stories about the TSA results. For example, state means were compared, and students in one state were reported as knowing x percent more math than those in another state. States with the lowest and highest means were highlighted, and the reactions of policymakers were included to justify their state rankings. Public reports should include simple graphs to highlight the variability, overlap, and uniqueness in states' score distributions.

2. *Apparent differences in student performance do not always reflect real differences.* Although the press briefing materials prepared by NCES and ETS specifically addressed this issue, the media reports ignored it and provided tables of comparative state ranks. Interval estimates of states' mean performances should be included in the public reports of NAEP TSA results, not point estimates of means and proportions.

3. *The press, the public, and those with policymaking authority will likely overgeneralize NAEP TSA results, even if warned not to do so.* For example, press reports suggested that "most students cannot meet on-the-job demands for problem solving" and "boys score higher across the board." As a countermeasure, the public reports of NAEP TSA results should include specific examples of interpretations that can and cannot be supported by results.

4. *Means do not tell the whole story about students' NAEP TSA performances.* Although the press release materials emphasized mean performances, substantial information about other characteristics of the distributions of performance was included. The investigator used some of these statistics to correlate state ranks at other percentiles and found wide differences in rank order. Several alternative statistics should be used to show state comparisons.

5. *Aggregated test results often mask as much as they describe.* The performances of subpopulations among eighth grade populations were not apparent, for example, among students in various racial or ethnic groups or students enrolled in specific mathematics courses. The results for subpopulations were analyzed, and the investigator developed data displays to illustrate their advantages, such as indicating the strengths of relationships.

6. *When testing agencies do not provide interpretations of test results, the public and the press will create their own.* Occasionally, but not frequently, these interpretations will be accurate. The data displays provided in the press release materials included correlates of between-state differences in average performance, but most reports ignored them, as well as press release statements about those differences. Graphical displays, as illustrated in examples, might convey that information more effectively.

7. *Reports of NAEP TSA results should not depend solely on the normative properties of the NAEP scale.* The descriptions of anchor levels provided to the press were too brief and sterile and did not accurately reflect the diversity of skills they represent. Many reports assumed that all students at eighth grade level should attain the maximum scores; none looked at score differences that might be educationally significant, rather than statistically significant. Within-grade content-area scales could be considered, and other strategies should be developed to identify and describe the educational significance of states' means and other distributional characteristics.

Reports for Policymakers

The reports intended to communicate NAEP TSA results to state-level decisionmakers were factual and well organized. After applying the reporting standards to the reports for state-level policy, the investigator found that most were met. Standard errors of measurement were reported, and scales and their rationales were clearly described. It is questionable whether the scales were consistent with the intended purpose; attention should have been paid to strengths of relationships. Follow-up studies should assess the utility of these reports.

The Case for District- and School-Level Results from NAEP

Ramsay Selden
Council of Chief State School Officers

The state-level NAEP has been implemented over the past five years with a good deal of consideration. Part of the political solution of how to implement state-level NAEP, however, was to put school- or district-level use of NAEP in limbo. Consequently, although the state-level program is being tried and evaluated, the question of whether NAEP should be used at the school or district level has not been answered. Several states have lobbied for it. On the other hand, most organizations representing local school systems or practitioners oppose it. In the absence of hard data based on empirical trial, this paper attempts to explore the arguments *for* using NAEP at the local level. In addition, some of the potential problems in using NAEP at this level are acknowledged and considered. Finally, to the extent possible, these potential problems are weighed in relation to the prospective benefits. The paper does not explore using NAEP to produce scores for individual students, a profoundly important concept that warrants careful exploration.

Arguments for School- and District-level NAEP

1. *Having information is better than not having it.* Most of the time, information is useful, and the more of it we have, the better, as long as the information is organized and presented in a way that renders it useful.

2. *Important, useful information can be generated through NAEP at the school and district level.* Student achievement data from NAEP is close to the U.S. state-of-the-art in most subjects. Subject-matter frameworks represent professional and political consensus, and they incorporate the state-of-the-art in theoretical thinking about performance in the subjects. As a professional consensus, NAEP has, in several subjects, reflected the resolution of major professional differences in its approaches to measuring performance outcomes. This is true in mathematics (where the assessment has been planned with consideration of NCTM *Standards*), reading (where the framework was the crucible for resolving important professional differences), and writing (where the assessment has changed to reflect substantial developments in the ways educators view the writing process).

In addition to student performance data, NAEP offers data on student background variables, teacher characteristics, instructional practices, and school management approaches. Few local assessment programs include these additional kinds of information, and for local schools or districts to gear up to develop this information on their own would be difficult, expensive, and wasteful.

NAEP also offers the tremendous benefit of comparable data from numerous sites. The value of this comparability is obvious for student performance data, but it is equally important for background data. Educational planning and management in the United States are done in an information vacuum—virtually in the absence of program data from large-scale data collection. Test scores, fiscal data, and staff evaluations are used, but data on programs is scarce. The potential for this kind of information to support analysis, planning, and management at the local level is tremendous.

3. *Educational data in the U.S. should be vertically integrated where possible—that is, similar at the local, state, and national levels.* The current U.S. system is balkanized and redundant. We must structure relationships among testing programs so that when a student is tested once, the results feed a number of purposes up and down the system. This would eliminate two or three layers of testing at least. In addition, results reported to educators and the public would be expressed in the same metrics and formats so that we could begin talking about testing in a single, common set of terms.

4. *We are arriving at national consensus on the goals and content of education.* A serious reason for not using NAEP at the local level was that it would lead to a national curriculum. This reason is valid only if a national curriculum does not exist and is not desired. As this is written, a National Council on Education Standards and Testing is being convened to study the feasibility of national standards for education. This move reflects a political consensus that some national focus of attention on the content of, and standards for, instruction is desired.

Aside from the argument as to whether we *ought* to have a somewhat centralized curriculum, there is the fact that relatively little diversity *exists*. Due to market forces in educational materials, testing, and teacher training, the content of instruction is not only low-level, fragmentary, and content-thin, it is consistently so.

Liabilities in a School- and District-level NAEP

1. *Use of NAEP in this way could discourage state and local innovation and development in assessment.* Assessment programs are costly, labor-intensive, and time-consuming to develop. Much depends on how local NAEP would be achieved. If linking and equating were used to achieve local reporting of NAEP scores, state and local programs could continue to be distinctive, while tying into the NAEP system. If NAEP were packaged and extended homogeneously, distinct local and state programs could be driven out as redundant luxuries.

2. *If NAEP is used as a high-stakes accountability device at the local level, this could lead to teaching to the test and distortion of the curriculum.* The main limitation to previous achievement tests, which resulted in their misuse, was the differences between the content of the tests and the content we wanted taught. Standardized, multiple-choice achievement tests have overemphasized relatively low-level operations, which are amenable to short-answer test item formats. Then, when the pressure is put on, it is natural for educators to focus on those narrow aspects of the curriculum, on which they know the tests depend, to bring up scores. Tests that would avoid these problems and misuses would be tests that fully represent the

curriculum and that measure important knowledge and capability in a valid and not immediately predictable way.

3. *Use of NAEP at the local level could interfere with the national program.* In addition to factors already discussed, use of NAEP at the local level could affect the national program by making it more difficult to maintain item security. Another effect would be in the student and exercise sampling design. To represent schools or small districts in the NAEP grade levels, a different sampling and spiralling approach could be needed, with potentially significant effects on the scaling and analysis of results.

4. *The cost of a local NAEP could be high.* NAEP generally costs more on a per pupil basis than commercial, standardized tests. If local use of NAEP could supplant much current state and local testing, savings could be achieved.

Overall Assessment of Impact

Two of the factors to be weighed most carefully are the informational value of NAEP against its potential for misuse. The informational value offered by NAEP at the local level is very great since new forms of information—high quality student achievement data and instructional program background data—would be made available in a powerfully comparable form. On the debit side is the potential for misuse. To minimize this threat, local use of NAEP would have to be implemented in thoughtful accountability systems, where consequences are realistically and carefully considered. And, the testing program would have to continue farther in its current direction of becoming an assessment program that is difficult to teach to inappropriately.

The other liability is whether implementation of local NAEP would drive out good local and state testing programs. This depends heavily on whether development of local NAEP is done through linking and equating systems or through monolithic expansion. If linking became available and economically feasible, it could be expected that states would use it to maintain particular features of, and purposes for, their testing programs, while tying into NAEP and generating NAEP scores for schools and districts. It would appear to be in the interest of assessment for states to be encouraged and enabled to develop statewide systems which are distinctive and creative, while tying into a national assessment system that provides local schools and systems valuable data in a common national currency.

Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting Below the State Level

Walter Haney
George F. Madaus
Boston College

This paper summarizes and comments on the proposals of the National Assessment Governing Board (NAGB) to remove the prohibition against the use of NAEP tests below the state level and those in *America 2000* for an individual version of NAEP. It reviews some of the reasons why NAEP was originally organized to prevent some of the things now being undertaken or considered; discusses some of the practical problems that likely would arise in development of individual versions of NAEP instruments; and summarizes what might happen if NAEP is used to make high-stakes decisions about districts, schools, or individuals.

The NAGB Proposal

On November 17, 1990, NAGB reaffirmed its December 9, 1989, position that Congress should remove the prohibition against use of NAEP tests and data reporting below the state level. NAGB proposed that a local option be given to states, districts, and localities to use NAEP. In terms of the 1994-1996 NAEP, NAGB proposed that "NAEP tests and items should be available to the states, districts, schools or organizations of schools that contract and pay for them. Publication of individual student scores should continue to be prohibited." Although NAGB wishes that NAEP be used by states, districts and schools, it does not spell out how the results would be used. However, NAGB notes that at the school district and school building level where many important decisions are made, such data would provide helpful information to parents and others interested in school improvement. Thus, it appears that NAGB would, initially at least, leave the use question up to the state and local education agencies. The use of NAEP results with individuals is somewhat confusing. Although NAGB affirms the prohibition against publishing individual scores, it goes on to discuss how differences in motivation and possible preparation for NAEP by local education agencies would be a serious problem in comparing the results of students who took NAEP knowing they would get individual scores with those in the national and state samples who are assured NAEP cannot "count" for them. This could be read to imply that although individual scores cannot be published, they can be given to individuals, or that state and local education agencies could use results to make decisions about individuals as long as individual scores are not published.

The America 2000 Proposal for NAEP

America 2000, the Administration's strategy for national education reform, calls for new American Achievement Tests to be available by September 1993, but because they cannot be developed that quickly, the Department of Education will ask Congress to authorize the rapid deployment of individual versions of existing NAEP tests in reading, writing, and mathematics, at least. Three purposes often ascribed to tests and assessments are directly set out in the *America 2000* proposal for the new American Achievement Tests, and implicitly for individual versions of NAEP instruments: (1) providing information for accountability to people outside of schools; (2) providing information to teachers to help them improve instruction; and (3) helping students learn.

Accountability

School accountability is one of the least important functions for testing and assessment because schools are one of our society's social institutions most often held accountable to many parties on many grounds (e.g., report cards, parent meetings, school board elections, attendance reports, compliance with state and federal regulations, press reports). Public opinion polls indicate that many school characteristics other than test scores concern people most. In addition, most people outside the schools generally want only summary information, not individual information, about how a school is performing in comparison with other similar schools in broad curriculum areas such as English and math.

Information to Help Teachers Improve Instruction

Research shows that teachers and school administrators are not typically knowledgeable about testing technology, and most tests do not tell them anything they have not learned about their students from classroom contact. When externally mandated tests "drive" instruction, as a high-stakes NAEP test would do, both teaching and learning can suffer; in addition, teaching to a particular test invalidates the test results as indicators of more general learning.

A high-stakes NAEP could come to define the curriculum in terms of both what gets taught and how it gets taught, creating a de facto national curriculum. This contradicts the stated goal of giving schools and their leaders more freedom and professional authority to make important decisions. NAEP also fails to meet teachers' needs for assessments that provide diagnostic information for individuals and groups in individual classes in particular subjects, that provide rapid feedback on assessment results, and that are integrated with teaching and learning.

Information to Help Students Learn

Research on the elements of testing that help someone learn indicates that feedback that is relatively rapid and that provides guidance to or identification of correct answers is more instructionally effective than when learners are simply told their answers were right or wrong. This indicates that the sort of new NAEP for individuals is simply not conducive to student learning. The *America 2000* proposal for versions of NAEP to be reported below the level of the states boils down to the simple logic that no test, NAEP or any other instrument, can reasonably serve the range of purposes suggested.

Reasons Behind the Original Reporting Prohibitions on NAEP

The current experiment with state comparisons and proposals to assess individual students alters the original safeguards that made NAEP possible and raises the arguments of NAEP's original critics that NAEP would be a national/federal test. If that happens, and if important rewards or sanctions are associated with performance, then teachers will teach to the NAEP exercises and subject frameworks, which, in turn, will become the curriculum of the nation.

NAEP was initially funded by private foundations but is now entirely funded by the federal government. A federal test aimed at individual students or schools under the auspices of NAEP raises central questions that have not yet been answered: Exactly what will the results be used for? Will they be used to certify, classify, sort, or select students—to make important decisions about them? Concern about the likely ill effects of a national high-stakes test appear to be shared by many professional colleagues. Before American students, already the most heavily tested in the world, are subjected to yet another testing treatment, the body of knowledge on the effects of testing should be evaluated and applied.

It would be wise to ask what subjects, content, and levels would be included in a national curriculum and how closely they align with the NAEP framework; what the adjective *national* means (commercial, private, quasi-governmental, or governmental); what such tests would be used for; what subjects/content/skills/abilities the test would and would not measure and at what grade(s) the tests would be administered; and who would sponsor, control, operate, and fund a national test and national curriculum.

Practical Problems

Since its beginning, NAEP has been designed not to test individuals but to produce national and regional results as efficiently as possible, with schools and students within schools sampled because there was no need to assess all students. NAEP

produces five "plausible values" for each student instead of individual student scores. These plausible values reflect the uncertainty of individual results and are useful for estimating the national and regional population proficiencies that NAEP reports. They do not, in general, meet the professional standards for individual test scores and are not appropriate for individual decisionmaking. NAEP's broad area coverage is appropriate for a wide spectrum of students but is not necessarily efficient or appropriate for detailed student reporting. A more targeted examination is necessary to allow individual students to show what they know and can do. Cost is another factor: NAEP's cost per student is reasonably small when compared to the alternative of testing every American student. Although costs such as test development and sampling could be reduced somewhat, much of the NAEP cost is simply administering and scoring the assessment, which would not be reduced. With the use of non-multiple choice items, problems arise concerning the cost, the time needed to score results, and the utility of individual reports.

One of NAEP's great strengths is its national norms, which are probably the best national educational norms ever assembled. If assessment conditions are changed, the performance of students can also be expected to change. If NAEP were to become a high-stakes test with consequences for individual students, then the NAEP norms would not be inappropriate for comparisons. Furthermore, to be useful, results would need to be reported quickly. Reporting of NAEP results has been delayed for technical analyses and for preparation of responsible interpretive reports, but delays of a year or more would largely vitiate the utility of results for current school-level accountability or for monitoring individual student progress.

The Problem of High-stakes Testing

The major conclusions from considerable literature on the impact of high-stakes testing on education can be summarized in six principles: (1) The more any quantitative social indicator is used for social decisionmaking, the more culpable it will be for distorting and corrupting the social processes it is intended to monitor. (2) The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon—if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false—the effect is produced by what individuals perceive to be the case. (3) If important decisions are presumed to be related to test results, then teachers will teach to the test. (4) In every setting where a high-stakes test operates, a tradition of past exams has developed which eventually comes to define de facto the curriculum. (5) When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement. (6) A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam.

If the prohibitions are lifted against the use of NAEP tests and data reporting below the state level, NAEP will become a high-stakes test with the results described in these principles. The conclusion is that many of the features of NAEP that have led to its technical excellence in reporting on national educational progress over time mitigate against its utility for such very different reasons.

Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot Be Made

Edward H. Haertel
Stanford University

This paper reviews some elements of the educational policy context that led to the Trial State Assessments, the structure and technical characteristics of the data collected, and some of the kinds of inferences, valid and invalid, that may be attempted.

Overview of the TSA Data

The 1990 and 1992 Trial State Assessments (TSAs) may initiate the most significant changes in the National Assessment of Educational Progress (NAEP) since its inception over 20 years ago. Enormous effort has been expended over a remarkably short period of time to fashion their content and procedures, create an enormous field organization, administer and monitor the data collection, and now to process, report, and interpret the data collected. It may at first seem surprising, in the light of all this effort, that the purposes the TSAs are to serve are not entirely clear.

Although TSA data have a few potentially important technical limitations, by and large, the TSA is close to the state-of-the-art in large-scale assessments. Even the lowest school-level and student-level participation rates and even the highest student-exclusion rates are within acceptable ranges. Comparability across states may be somewhat threatened by variations across states in the degree of cooperation obtained, but it is very unlikely that sampling or other technical problems will lead to any gross distortions of estimated achievement levels.

The limited content of the TSA exercise pools is a more serious concern to the investigator. The consensus process leading to the assessment frameworks were implemented and executed conscientiously, and exercise development was highly competent. Nonetheless, the tension between assessing achievement as it is versus as it ought to be, the severe time constraints under which the exercise pools were developed, and both economic and logistical constraints on exercise timing and scoring resulted in an exercise pool for mathematics that some experts have found disappointing. Similar factors seem likely, in the end, to temper the boldness of the 1992 reading framework. In eighth grade mathematics, the TSA exercise pool is even more limited due to the exclusion of three blocks of so-called "higher-order" and "estimation" exercises. Ultimately, the TSAs can directly inform only those learning outcomes actually assessed. The 1990 TSA will provide a limited view of achievement in mostly conventional mathematics at one grade level, in about two thirds of the states. The 1992 TSA will provide about the same kind of information in eighth grade mathematics; the fourth grade reading and mathematics exercise pools are yet to be developed.

Valid and Invalid Inferences

Inferences from the TSA data have both a statistical aspect and a substantive aspect. There may be a tendency to dwell more on the problems of valid statistical inference from a sample to a population than on the substantive assumptions required to support interpretations of the meaning of TSA findings or their implications for action. Nonetheless, the substantive part of the inference is by far the weaker. Interpretations of TSA findings may be especially prone to problems of faulty causal attribution and of overgeneralization beyond the content actually assessed.

The most important and most defensible inferences from the TSA data may well be those for which the substantive assumptions are weakest, simply documenting achievement distributions at the state level and using selected exercises from the NAEP exercise pool itself to illustrate the meaning of selected score levels. It will be especially important to document achievement disparities among subpopulations, which imply inequities in our provision of a sound education to all of our nation's children. This is no more than an extension of the historical role of NAEP to the state level. The states are responsible for public education, and for that reason, such information may have greater impact when reported at the state level than for the nation as a whole.

The 1990 and 1992 TSA data will not be useful for describing achievement in the nation as a whole, and even if all 50 states eventually participate, the use of the pooled TSA data to supplant the national NAEP sample may be unwise. The TSAs will also be of quite limited usefulness in characterizing the overall quality of states' educational systems or the effects of particular policies, reforms, or instructional practices, primarily due to limitations of the exercise pool and limitations on the identification of data below the state level with particular schools or districts. (Haertel cautions that his comments about limitations arising from the prohibition against reporting TSA results below the state level should not be construed as advocating such reporting. The question of reporting below the state level is quite complex and is beyond the scope of his paper.) In principle, two types of allocation decisions might be informed by the TSA data: allocation of instructional time and other resources to one versus another learning objective and allocation of instructional resources to one versus another area or subpopulation. Due to the limited potential of the TSA exercise pools to differentiate among learning outcomes, the data will be of limited usefulness for the first type of allocation decision, but they may have implications for the second type of allocation decision. Finally, the TSA data may be of occasional value in addressing more focussed questions relating to the specific areas assessed. Such analyses will be somewhat opportunistic and will depend heavily on the scope and quality of the TSA background data. The usefulness of the TSA data for such purposes will increase significantly over the years as trend lines are established.

References

Alexander, Lamar and James, H. Thomas. *The Nation's Report Card: Improving the Assessment of Student Achievement*. Washington, D.C.: National Academy of Education, 1987.

American College Testing Program. *Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading and Writing*. Iowa City, IA: 1991

American Institutes for Research. *National Dropout Statistics Field Test Evaluation: Final Report*. Palo Alto, CA: 1991.

Angoff, William. "Scales, Norms and Equivalent Scores," in *Educational Measurement*. Ed. R. L. Thorndike. Washington, D.C.: American Council of Education, 1971.

Beaton, Albert et al. *Implementing the New Design: The NAEP 1983-84 Technical Report*. Washington, D.C.: National Center for Education Statistics, 1987.

Beaton, Albert et al. *Expanding the New Design: The NAEP 1985-86 Technical Report*. Washington, D.C.: National Center for Education Statistics, 1988.

Bourque, Mary Lyn and Garrison, Howard. *The Levels of Mathematics Achievement, Initial Performance Standards for the 1990 NAEP Mathematics Assessment, Volume 1: National and State Summaries*. Washington, D.C.: National Assessment Governing Board, 1991.

Chira, Susan. "Major Panel on Education Urges Nationwide Standards and Tests." *New York Times* (December 18, 1991): A-1

Cooper, Kenneth J. "Education Goals Called Far from Realization: Report will serve as Baseline for Progress." *Washington Post* (October 1, 1991): A-4.

Educational Testing Service. *Mathematics Objectives: 1990 Assessment*. Princeton, NJ: 1988.

Ekstrom, Ruth; Goetz, Margaret; Pollack, Judith; and Rock, Donald. "Who Drops Out of High School and Why? Findings From a National Study," in *School Dropouts: Patterns and Policies*. Ed. G. Natriello. New York: Teachers College, Columbia University, 1987.

Haertel, Edward. "Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot Be Made," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Haney, Walter and Madaus, George F. "Cautions on Removing the Prohibition Against the Use of NAEP Tests and Data Reporting Below the State Level," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Jaeger, Richard. "General Issues in Reporting of NAEP Trial State Assessment Results," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Koffler, Stephen et al. *The Technical Report of NAEP's 1990 Trial State Assessment Program*. Washington, D.C.: National Center for Education Statistics, 1991.

Koretz, Daniel; Linn, Robert; Dunbar, Steve; and Shepard, Lorrie. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests," in *Annual Meeting of the American Research Association and the National Council on Measurement in Education*. Ed. D. Koretz. Chicago: Rand Corporation, 1991.

Linn, Robert; Shepard, Lorrie; and Hartka, Elizabeth. "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics and Subpopulation Breakdowns," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

McLaughlin, Donald H.; Stancavage, Frances B.; Chambers, Jay G.; Hartka, Elizabeth; and Ercikan, Kadriya. "Evaluation of the Administration of the 1990 Trial State Mathematics Assessment," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

National Academy of Education. "Review of the Report by a Committee of the National Academy of Education," in *The Nation's Report Card: Improving the Assessment of Student Achievement*. Washington, D.C.: 1987.

National Assessment Governing Board. *The Levels of Mathematics Achievement*. Washington, D.C.: Aspen Systems, 1991.

National Council on Education Standards and Testing. *Draft Discussion Report on Standards*. Washington, D.C.: October 16, 1991.

National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: 1989.

National Education Goals Panel. *The National Education Goals Report: Building a Nation of Learners*. Washington, D.C.: 1991.

Olson, L. "'Confusing' Array of Players Charts Course Toward National Standards." *Education Week* 11(8) (October 23, 1991): 1.

Selden, Ramsay. "The Case for District- and School-Level Results from NAEP," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Silver, Edward; Kenney, Patricia; and Salmon-Cox, Leslie. "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Spencer, Bruce. "A Critique of Sampling in the 1990 Trial State Assessment," *in Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford; CA: National Academy of Education, 1992.

Spencer, Bruce. "Eligibility/Exclusion Issues in the 1990 Trial State Assessment," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education, 1992.

Stancavage, Frances B.; Roeber, Edward; and Bohrnstedt, George. "Study of the Impact of Reporting the Results of the 1990 Trial State Assessment," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: National Academy of Education,1992.

Stufflebeam, Daniel; Jaeger, Robert; and Scriven, Michael. *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-1991 Effort to Set Achievement Levels on the National Assessment of Educational Progress*. Washington, D.C.. August 1991.

U.S. Department of Education. *America 2000: An Education Strategy*. Washington, D.C.:1990.

U.S. Department of Education, National Center for Education Statistics. *America's Challenge: Accelerating Academic Achievement*. Washington, D.C.: 1990.

U.S. Department of Education, National Center for Education Statistics, Mullis, Ina V. S. et al. *The State of Mathematics Achievement*. Washington, D.C.: 1991.

U.S. Department of Education, National Center for Education Statistics. *Dropout Rates in the United States*. Washington, D.C.: 1990.

U.S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills. *What Work Requires of Schools: A SCANS Report for America 2000*. Washington, D.C.: June 1991.

Westat Inc. *Report on Field Operations and Data Collection Activities, 1990 National and Trial State Assessments*. Rockville, MD: 1990.

Westat Inc. *Summary for All States of Training and Data Collection Activities, Preliminary Results, 1990 Trial State Assessment*. Rockville, MD: 1990.

Acronyms Appearing in this Report

CCSSO	Council of Chief State School Officers
ETS	Educational Testing Service
IEP	Individualized Education Plan
LEP	Limited English Proficiency
MSEB	Mathematical Sciences Education Board
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NALS	National Adult Literacy Study
NCES	National Center for Education Statistics
NCEST	National Council for Education Standards and Testing
NCTM	National Council for Teachers of Mathematics
NEGP	National Education Goals Panel
NESAC	National Education Standards and Assessment Council
SCANS	Secretary's Commission on Achieving Necessary Skills
TSA	Trial State Assessment